UNIVERSITY OF CALIFORNIA
SANTA CRUZ

REEF FISH POPULATION GENOMICS AND HYBRIDIZATION USING RADSEQ: A CASE STUDY WITH *DASCYLLUS TRIMACULATUS*

A dissertation submitted in partial satisfaction
of the requirements of the degree of

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Eva M. Salas De la Fuente

December 2016

The Dissertation of Eva M. Salas De la Fuente
is approved:

Professor Giacomo Bernardi, Chair

Professor Luiz A. Rocha

Professor Mark H. Carr

Professor Michael L. Berumen

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

List of Figures

Chapter 1

Chapter 2

Chapter 3

Abstract

REEF FISH POPULATION GENOMICS AND HYBRIDIZATION USING RADSEQ: A CASE STUDY WITH *DASCYLLUS TRIMACULATUS*

Eva M. Salas De la Fuente

The ecological and evolutionary dynamics of populations vary at different spatial and temporal scales. This is especially evident in the case of reef fishes, with bi-partite life cycles where the adult stage is benthic and sedentary, while its larval stage is capable of traveling long distances in the plankton. Understanding the genetics of populations at different evolutionary and ecological scales can help reveal metapopulation dynamics and speciation in the sea. Here I investigate population genomic patterns of the reef fish *Dascyllus trimaculatus* within a biogeographic province (the Red Sea and Arabian Peninsula), within an ocean basin (the Indian Ocean, that includes two biogeographic provinces), and at the intersection of biogeographic regional boundaries (at Cocos-Keeling Islands and Christmas Island, where Indian and Pacific Ocean marine faunas overlap). The goal is to understand the past and contemporary patterns of dispersal, and explore how relevant small and large-scale processes are to population subdivision and speciation. I used Restriction Site Associated DNA (RAD) markers to generate SNPs to study population genetic structure. In Chapter 1, I found strong genetic differentiation within the Red Sea. The region of Aqaba and the Northern Red Sea are different from the Central Red Sea and there is a genetic break between the Central Red Sea and the Farasan Islands. I also found differences between the Red Sea and the Arabian Peninsula populations. Most of the genetic structure was found in the loci outliers, suggesting that divergent selection may be acting on the species and allowing it to adapt to different environmental conditions. In Chapter 2, I found genetic differentiation of *D. trimaculatus* populations between the Nortwestern Indian Ocean biogeographic province (Red Sea and Arabian Peninsula) and the Western Indian Ocean province (Indian Ocean islands and the African coastline), supported by analysis with neutral loci and outliers. The differences between provinces may be attributed to the complex history of the Red Sea and Arabian Peninsula, the lack of suitable habitat on the coasts of Oman and Somalia, strong upwellings and the changing environmental conditions found in the Red Sea and Arabian Peninsula. The Indian Ocean *D. trimaculatus* has strong genetic differences with the Pacific Ocean populations. In Chapter 3, I studied *D. trimaculatus* populations of Pacific and Indian Ocean genetic clades in islands where the two groups overlap. I found cryptic hybridization between the Indian Ocean and Pacific Ocean *D. trimaculatus* genetic clades in the marine suture zone of Cocos-Keeling Islands and Christmas Island. I also found that there are consistent differences between the Pacific and Indian Ocean clades in the color of the rear of the dorsal fin. Future work will separate the Pacific and Indian Ocean *D. trimaculatus* into two species. This body of work shows the powerful approach of SNP markers to detect cryptic patterns of population structure, which remain difficult to discover using traditional genetic markers.

Dedication

To Priscilla Zamora-Trejos and Ana Cecilia Fonseca-Escalante: Marine biology was really fun with you, I wish you were here to write more projects together! To abuelito, Luis Angel Salas-Fonseca: Somewhere in the family there was another scientist! You showed me the beauty of nature and butterflies. To abuelita Grace Munoz de Salas, nana Cristina Padilla De la Fuente and Tía Carmen, who left while I was doing my PhD.

I also want to dedicate my dissertation to all of those that are still here: Specially to the love of my life Jagjit Chadha, my mom and dad, Milagro De la Fuente and Luis Guillermo Salas, my brother Luis Diego Salas and his wife Carolina Mora, to my nieces Irene and Alana Mora, and to my "sister" Augusta Anderson. This journey would have been impossible without all of you!

Acknowledgements

Thank you Jagjit! Thanks for understanding the moments of thesis craziness. You reminded me to take care when I forgot about myself. And thanks for helping me with the design of many figures that are part of this thesis, coadvisor! Thanks to my big family in Costa Rica: I moved out of the country to follow my dreams. I haven't been there for many important moments, but you always receive me with the open hearts. You let me be free and not everyone has that opportunity. Thanks for all the support! Thanks in special to mom and dad.

I want to thank my advisor Luiz Rocha. He's taught me how to be more efficient and pushed me to make my science relevant to a wider scientific audience. He has the skill to see an important pattern behind what appears dull and uninteresting. I am slowly learning that, but I have a long way to go! I also want to thank you for the opportunity to transfer my PhD to UCSC. Thank you Luiz, that decision was very important to me too, and it changed my life in great ways. All of this would have been impossible without your support, and the funding of the Lakeside Foundation of California Academy of Sciences, and later on the support of University of California Santa Cruz: for that, I want to thank Giacomo Bernardi, Pete Raimondi, Tyrus Miller, Kris West and Jim Moore for finding financial support at the end of my PhD.

There is not enough paper on this dissertation to thank Giacomo Bernardi! Thanks for your availability, your friendship, and for being an amazing mentor in science and life. Also, thank you for the many, many, many samples, which are the backbone of this dissertation. Thanks for the rest of my committee: Michael Berumen and Mark Carr! Really, thanks for all my committee for being always available, supportive and for making the journey really fun! Mike: thanks for the opportunity to do my work at KAUST and receive me like your lab student, with a desk, a lab, an abaya and even a bike, the most important thing to have in KAUST! I did 90% of my thesis work in KAUST. Thanks for the opportunity to do my fieldwork there and use the lab.

I've been lucky enough to belong to four labs. At Calacademy, nothing would have been possible without the CCG lab and the Ichthyology department. Brian Simison and Joe Russack maintain the core computer of data analysis. But beyond that, they are always available to help with everyday issues with our large datasets. Brian set up a system at Calacademy that is the foundation to all my work (and anyone that does genetics at Calacademy). Anna Sellas maintains the cleanest and most organized lab I ever seen, along with my new labmate, Cerise. Cerise, how many times you saved my life, thank you!! At the Ichthyology department, I want to thank Mysi Hoang, Dave Catania, Jon Fong and Claudia Rocha. Claudia was extremely patient with me and oriented me in the CCG lab and shared all her protocols that worked perfectly. Thanks to all my labmates: Michelle Gaither, Moisés Bernal, Hudson Pinheiro, Iria Fernandez. Thanks to my labmates of Giacomo's lab: Eric García, Gary Longo, Jimmy O'Donell, Kim Tengarddjaja, Alexis Jackson. Thanks to my labmates at KAUST: specially to Remy Gatins, May Roberts, Alex Kattan, Pablo Saenz, Joey Di Battista, Vanessa Robitzch, Maha Khalili, Marcela Herrera, Veronica Chaldez, Diego Lozano, Song

He, Lu Song, and my dear friend Celia Schunter that came to me to explore KAUST for the first time! Eternal thanks to Jean Paul Hobbs, thanks to him I got the most exciting chapter of my thesis. Thanks to Nicole Crane. She's taught me alternative academic paths that a researcher can take. Thanks to my friends Angela Quiros, Elsie Tandjaja and Augusta Anderson, for being there so many times!

I also want to thank some people and funding agencies for the specific chapters of this dissertation. For Chapter 1, I'd like to thank Joey Di Battista, Pablo Saenz Agudelo, Vanessa Robitzch, Remy Gatins, May Roberts, Song He, Diego Lozano, Alex Kattan. For Chapter 2, I would like to thank Jean Paul Hobbs, Joey DiBattista, Claudia Rocha, Anna Sellas, Brian Bowen, Moises Bernal, Hudson Pinheiro, Iria Fernandez-Silva, Shelley Jones, D. Wagner, Joshua Copus and Brian Simison for help in the laboratory, sample collections and/or fruitful discussions. Thanks to Jagjit Chadha and Angela Quiros for their timely contributions. I would also like to thank the British Indian Ocean Territory Administration, the Chagos Conservation Trust, and Charles and Anne Sheppard for obtaining permits and facilitating the expedition on which materials were collected from the Chagos Archipelago. For the Chapter 3, I also want to thank people that helped surveying the reefs. Tane Synclair-Taylor and Karen Wilshaw for pictures from Cocos and Christmas. Elsie Tandjaja, Tane Synclair-Taylor, Joey Di Battista, for samples and field observations. Thanks to Iria Fernandez and Carlos Prada for insightful discussions. Thanks to Jagjit Chadha for help with the figures of Chapter 3.

The work on my dissertation was possible with funding from California Academy of Sciences, a Lakeside Foundation scholarship, a Myers trust grant, funding from University of California Santa Cruz (UCSC) and funding from the King Abdullah University of Science and Technology (KAUST).

Introduction

Coral reefs are highly diverse and complex ecosystems, which provide shelter to thousands of species, and food and economic revenue for humans. Reefs can be patchily distributed or continuous, forming extensive barriers comprised of lagoons, crests and walls, or alternatively small and fragmented fringing reefs or coral patches. To add to the spatial complexity, these are widely distributed along the tropics, from highly continuous continental shelves to isolated oceanic islands.

One of the most important drivers of evolution of coral reef-associated species is the degree of connectivity between reef–associated populations. The majority of marine organisms reproduce by releasing thousands of gametes, which turn into tiny larvae floating in the open ocean. Imagine the life of larvae: where do the currents carry them? And how do larvae connect all these reef-associated populations? How does population connectivity via larval dispersal influence the patterns and rate of evolution, including speciation, of coral reef species? Most importantly, how do the different spatial scales of dispersal and population connectivity found in coral reefs affect the ecology and evolution of its inhabitants?

The importance of spatial scale in ecology and evolution is widely recognized. Our interpretations of biological processes can dramatically change depending on the scale at which the pattern is described. In population genetics, processes that influence gene flow may only be identifiable at certain scales.

Patterns of gene flow reveal the footprint of larval dispersal through thousands or millions of years, and also recent dispersal events.

The genome of any organism can be thought of as a record of millions of years of history. Genes with faster rates of evolution manifest recent history; while genes that evolve slowly reflect older processes (Wang 2010). The field of phylogeography has developed powerful coalescent analysis to detect deep and shallow branching events. When the branches are shallow, landscape and population genetics become most informative (Wang 2010). The latter field is especially good at detecting contemporary patterns of dispersal, if appropriate genetic markers are used (Wang 2010).

Recent studies have shown how phylogeographic analysis and landscape genetics can be an excellent complement to each other (Braaker & Heckel 2009; Pease *et al.* 2009; Wang 2010). A combined approach can help disentangle the force of contemporary versus historical events on a species (Zellmer & Knowles 2009). Mathematical models of species' range shifts often require the integration of phenomena that occur at multiple scales (Levin 1992), because minor differences in the local dynamics of individuals could result in radically different patterns at larger scales (Chave 2013). These models are necessary to understand the consequences of climate change, habitat loss and the design of protected areas (Levin 1992), but could not be accomplished without information about the processes acting across multiple spatial scales. Future work could also identify predictable relationships between these processes to inform these models.

My thesis is about the molecular ecology of a coral reef fish at different spatial scales. I am applying genomics to evolutionary and ecological questions, concerning the dispersal of marine larvae, the connectivity of populations between coral reefs, and evolutionary barriers to gene flow. My model species is the three-spot damselfish, *Dascyllus trimaculatus.* I use RADSeq to generate thousands of SNP markers to elucidate patterns of gene flow among populations.

Study species: *Dascyllus trimaculatus*

My study species is the three-spotted dascyllus, *Dascyllus trimaculatus* (Pomacentridae). It is a widespread, territorial damselfish that can live from zero to 55 meters deep, inside and outside of reef lagoons (Bernardi *et al.* 2001). *D. trimaculatus* is a demersal spawner. Males prepare and guard nests where females lay eggs (McCafferty *et al.* 2002). After 3-4 days the larvae hatch (Bernardi *et al.* 2001; Garnaud 1957). Its pelagic larval duration (PLD) is about 22-26 days (Wellington & Victor 1989). *D. trimaculatus* recruits settle into anemones and spend a few months as juveniles before migrating to the reef. Although settlers have been observed in habitats other than anemones, this is uncommon (Leray *et al.* 2010). Eventually, juveniles leave the dependency with anemones, but as adults they tend to stay nearby. *D. trimaculatus* feeds on plankton and forms large feeding groups over the reefs. Post-settlement mortality is high, but the juveniles that survive in the anemones have greater survivorship in the transition from the anemone to coral reefs (Bernardi, comm. pers.). *D. trimaculatus* was chosen because it is small, easy to collect and study, much genetic work has been done on it, it has a short PLD, and it has a widespread distribution, encompassing the Red Sea, Indian Ocean and extends to the Central Pacific Islands.

This species belongs to a complex divided into five major mitochondrial clades: 1) the Marquesas endemic *D. strasburgi*, 2) the Hawaiian *D. albisella* with cryptic divergence, 3) the French Polynesian *D. trimaculatus,* 4) the Pacific rim clade comprising two introgressed groups: *D. trimaculatus* and *D. auripinnis*, and 5) the Indian Ocean *D. trimaculatus* (Bernardi *et al.* 2003; Bernardi *et al.* 2001; Bernardi *et al.* 2002; Leray *et al.* 2010; McCafferty *et al.* 2002). Earlier studies described a well-mixed Indian Ocean group, based on mitochondrial control region and microsatellite data (Bernardi *et al.* 2002; Leray *et al.* 2010). The Indian Ocean population was also reported as the *D. trimaculatus* basal group and exhibited the highest genetic diversity (Leray *et al.* 2010).

Study system: the Indian Ocean

The Indian Ocean is located between Africa, Asia and Australia. It is a ~8,000km wide basin that extends from East Africa to Western Australia, and contains about 3,000 islands with mountains and hills and coral islands and atolls. The Indian Ocean is one of the warmest bodies of seawater. Its weather is influenced by monsoons and seasonal cyclones. The prevailing currents flow clockwise in the northern hemisphere with reversals during the winter, and flow and counterclockwise in the southern hemisphere (Allen & Steene 1987).

The three chapters of this dissertation provide insight into the underlying processes that shape and maintain the marine biodiversity of coral reef fishes within the Arabian Peninsula and the Indo-Pacific Ocean. Chapter one is about population connectivity of *Dascyllus trimaculatus* in the Red Sea and Arabian Peninsula. The first chapter's goal is to investigate the population connectivity of these two regions, and the fine scale patterns of genetic structure within the Red Sea. Specifically I am interested in evaluating the effect of a potential dispersal barrier at 20 degrees N of latitude within the Red Sea. In this chapter I discovered that this species shares the same genetic breaks as other species found in the region. Interestingly, it only shows these patterns in outlier loci that are candidates for divergent selection.

In Chapter two, I explore population genomics of *Dascyllus trimaculatus* at a larger scale within the Indian Ocean. The objective is to assess population structure in Indian Ocean *D. trimaculatus* and to consider the relative contribution of neutral and adaptive processes in partitioning diversity in this geographic region. I hypothesize that Indian Ocean *D. trimaculatus* are not a single panmictic population, and that increased resolution will reveal genetic differentiation across the region. In this chapter I find genetic divergence between Arabian Peninsula and Western Indian Ocean populations, in both neutral and outlier loci, but the pattern is clear-cut in the outlier loci. Certain portions of the genome reveal that coral reef fishes, despite their capacity of long distance dispersal, show a strong signal of local retention or alternatively, local adaptation.

In Chapter three, I zoom out to a region that is the crossroad between Indian and Pacific Ocean fauna. This chapter is about the genomic structure of *Dascyllus trimaculatus* in a marine suture zone, where hybrids of many reef fishes have been found. Data from chapter three shows that the Pacific and Indian Ocean clades of *Dascyllus trimaculatus* hybridize at Cocos-Keeling Islands and Christmas Island, and that the dynamics of hybridization are different in each island, leading to different evolutionary outcomes.

This dissertation will help advance the science needed for conservation in the Indian Ocean, the Arabian Peninsula and highlight the importance of Cocos-Keeling Islands and Christmas Island as regions of evolutionary importance. I had the opportunity to study reefs very difficult to access. I also had the chance to meet many collaborators doing research in the Red Sea and Indian Ocean. I am working with several scientists from the region and I want to make these data available to help identify key conservation units in the Indian Ocean, as well as to help build management plans in Saudi Arabia.

References

Allen G, Steene R (1987) *Reef fishes of the Indian Ocean* T. F. H. Publications, Neptune City, NJ.

Bernardi G, Holbrook S, Schmitt R, Crane N (2003) Genetic evidence for two distinct clades in a French Polynesian population of the coral reef three-spot damselfish *Dascyllus trimaculatus*. *Marine Biology* **143**, 485-490.

Bernardi G, Holbrook SJ, Schmitt RJ (2001) Gene flow at three spatial scales in a coral reef fish, the three-spot dascyllus, *Dascyllus trimaculatus*. *Marine Biology* **138**, 457-465.

Bernardi G, Holbrook SJ, Schmitt RJ, Crane NL, DeMartini E (2002) Species boundaries, populations and colour morphs in the coral reef three–spot damselfish (*Dascyllus trimaculatus*) species complex. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **269**, 599-605.

Braaker S, Heckel G (2009) Transalpine colonisation and partial phylogeographic erosion by dispersal in the common vole (Microtus arvalis). *Molecular Ecology* **18**, 2518-2531.

Chave J (2013) The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecology Letters* **16**, 4-16.

Garnaud J (1957) Ethologie de *Dascyllus trimaculatus* (Rüppell). *Bull. Inst. Oceangr. Monaco.* **54**, 1-10.

Leray M, Beldade R, Holbrook S, Schmitt R, Planes S, Bernardi G (2010) Allopatric divergence and speciation in a coral reef fish: The three-spot dascyllus, *Dascyllus trimaculatus* species complex. *Evolution* **64**, 1218-1230.

Levin SA (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* **73**, 1943-1967.

McCafferty S, Bermingham E, Quenouille B, Planes S, Hoelzer G, Asoh K (2002) Historical biogeography and molecular systematics of the Indo-Pacific genus Dascyllus (Teleostei: Pomacentridae). *Molecular Ecology* **11**, 1377-1392.

Pease KM, Freedman AH, Pollinger JP, McCormack JE, Buermann W, Rodzen J, Banks J, Meredith E, Bleich VC, Schaefer RJ (2009) Landscape genetics of California mule deer (*Odocoileus hemionus*): the roles of ecological and historical factors in generating differentiation. *Molecular Ecology* **18**, 1848-1862.

Wang IJ (2010) Recognizing the temporal distinctions between landscape genetics and phylogeography. *Molecular Ecology* **19**, 2605-2608.

Wellington G, Victor B (1989) Planktonic larval duration of one hundred species of Pacific and Atlantic damselfishes (Pomacentridae). *Marine Biology* **101**, 557-567.

Zellmer A, Knowles LL (2009) Disentangling the effects of historic vs. contemporary landscape structure on population genetic divergence. *Molecular Ecology* **18**, 3593-3602.

Chapter 1

Connectivity of *Dascyllus trimaculatus* in the Arabian Peninsula

CHAPTER 1

CONNECTIVITY OF *DASCYLLUS TRIMACULATUS* IN THE ARABIAN PENINSULA AND THE RED SEA

ABSTRACT

The study of connectivity between populations is key for the design of Marine Protected Areas (MPAs), and it is important to include multiple species to account for the variability of life histories. Here, I contribute with a population genomic study of the three-spot dascyllus (*Dascyllus trimaculatus*), a common and widespread reef fish throughout the Indo-Pacific. The aim of my study was to determine the patterns of genetic structure in the Red Sea and Arabian Peninsula and to test if this study species shows a genetic break within the Red Sea associated with environmental changes found south of 20°N. I sampled populations within the Red Sea along the coast of Saudi Arabia, and outside of the Red Sea in the Arabian Peninsula, and carried out analysis using SNPs generated by RADSeq, using neutral and outlier loci. Neutral loci revealed low but significant genetic structure between the Red Sea and Arabian Peninsula, and low genetic structure within the Red Sea populations. Analysis of loci outliers under divergent selection show strong genetic differences between the Red Sea and Arabian Peninsula, in particular there is a strong difference between Socotra and the rest of populations. Within the Red Sea, outliers revealed a strong genetic break at 20°N, coarsely concordant with the previously identified genetic break but not exactly in the same location previously found for low-dispersal species. The patterns of genetic differentiation found in the outliers may be due to

adaptation to the steep environmental gradient in the Red Sea, and future studies should investigate what mechanisms and what genes are involved in these changes.

INTRODUCTION

Dispersal is an important feature in marine populations; it can influence the distribution and range of species (Bowler & Benton 2005), the dynamics and persistence of populations, and from an evolutionary perspective, determine the levels of gene flow and local adaptation (Dieckmann *et al.* 1999). Dispersal from the parental habitat is advantageous as it can reduce intraspecific competition and the likelihood of inbreeding, and facilitate range expansion and diminish local extinction risk (Pechenik 1999). However, it can also lead to high mortalities during the dispersal stage, with no guarantee of finding suitable habitat (Gadgil 1971), and in the long term, may prevent adaptation to local conditions (Pechenik 1999). Thus it is unclear whether or not selection would favor recruitment in close proximity to parents.

Resolving dispersal abilities is an important task for understanding the ecology of populations, communities and ecosystems. It can help determine population boundaries or the existence of metapopulations (Weersing & Toonen 2009). With this knowledge, ecologists can improve the accuracy of predictions about population distribution and dynamics. This is also a matter of interest for the conservation of fish stocks and the design of marine protected areas (Botsford *et al.* 2009). Ideally, managers would like to identify source populations and protect them, so that they can replenish other populations. These important questions have formed part of the basis for the study of marine population connectivity.

In the field of population genetics, traditional measurements, such as assessing changes in allele frequencies, have been useful for determining population boundaries, particularly when they are strong and usually at large spatial scales (Hellberg 2009). They also have been useful for looking at isolation by distance and dispersal kernels (Hellberg 2009). One of the biggest issues is that traditional population genetic approaches and traditional markers often cannot resolve population boundaries in marine organisms, due to their large effective population sizes (Benestan *et al.* 2015). In such large populations, a few migrants per generation can conceal population structure at contemporary scales. For example, *Thalassoma lucansanum* was found to have high self-recruitment using otolith chemistry data (Swearer *et al.* 1999), however, it was also found to lack genetic differentiation across the whole Caribbean sea, based on analysis with microsatellite data (Haney *et al.* 2007; Purcell *et al.* 2006). This means traditional genetic markers may be missing the effect of important contemporary processes.

The best approaches to assess population connectivity are the use of parentage analysis and relatedness tests (Almany *et al.* 2007; Christie *et al.* 2010; Jones *et al.* 1999; Jones *et al.* 2005). Some of the limitations of these methods are that a large representation of the population needs to be sampled, limiting the number of possible studies due to cost and intensive fieldwork (especially for parentage analysis). A possible solution is to explore the effect of more recent processes with the use of high-resolution genetic markers. The extraction of SNPs with RAD Sequencing has become a popular approach, where now it is possible to obtain thousands of loci and classify them as loci under selection and neutral loci. Therefore, scientists have access to higher resolution due to the large number of markers, and also with the possibility to study the contemporary role of selection and adaptation.

Here, I contribute with a population genomic study of the three-spot dascyllus (*Dascyllus trimaculatus*), a common and widespread reef fish throughout the Indo-Pacific. The aim of this study is to determine the patterns of genetic structure in the Red Sea and Arabian Peninsula using a hierarchical approach. I first determine the patterns of genetic structure at a large scale including the whole study region, and then focus on the Red Sea. I analyze neutral and outlier loci to test if this study species shows a genetic break within the Red Sea associated with environmental changes found south of 20°N, and to test if there are genetic differences between the Red Sea and the Arabian Peninsula using RADSeq.

MATERIALS AND METHODS

Sample collections and data analysis

I obtained samples of *Dascyllus trimaculatus* during research expeditions between 2013 and 2015 (except Oman, in 2004). Individuals were collected using Hawaiian slings or hand nets. Immediately after collection, fin clips were placed in 95% ethanol and stored at -20° C in the laboratory. Samples were categorized into eight populations (Fig. 1.1), based on results of earlier connectivity studies on the Red Sea and Arabian Peninsula (Nanninga *et al.* 2014; Saenz-Agudelo *et al.* 2015).

Library preparation

Genomic DNA was extracted from preserved tissue samples. The NucleoSpin® 96 tissue kit (Macherey-Nagel) was used according to the manufacturer protocol. DNA concentration was measured in a SpectraMax® Plus 384 Microplate Reader (Molecular Devices), using the Qubit HS dsDNA essay kit (Life Technologies), and each sample was standardized to 500 ng. Two libraries with 96 individuals each were prepared using the double-digest RADSeq protocol described by Peterson *et al.* (2012), with modifications. The protocol uses two barcoding steps, one in the ligation and another one in the PCR. Samples were digested for 3 hours at 37°C, using restriction enzymes SphI and MluCl (New England Biolabs). Digests were quantified with Qubit HS dsDNA essay kit (Life Technologies) and a Qubit 2.0 Fluorometer, and then cleaned with Dynabeads M-270 Streptavidin (Life Technologies). A set of 16 unique barcodes was used during ligation. After ligation, the first sets of barcoded individuals were pooled and bead cleaned, to obtain six pools per library. Pools were size-selected for a range of 400 bp using a 2% agarose gel that was ran for 45 minutes, and purified with the Zymoclean^{TM} Gel DNA Recovery Kit (Zymo Research). Pools were barcoded using unique Illumina Indexes and amplified with 10 PCR cycles, using the high-fidelity Platinum Taq DNA polymerase (Thermo Fisher Scientific). The size selection of the pools was verified using a High Sensitivity Kit on a 2100 Bioanalyzer (Agilent Technologies). Product concentration was measured before pooling into single libraries. A qPCR was used for quantification using a KAPA Library Quantification Kit with an Illumina Primer premix kit, run in an ABI 7900HT fast real-time PCR system. Pools then were standardized and merged. Two libraries consisting of 6 uniquely barcoded pools each (96 individuals), were sequenced in two lanes on an Illumina Hi-Seq 2000, at KAUST Genomics Core facilities, which resulted in 292,778,914 single end reads altogether (101bp).

Loci assembly

Loci were assembled using STACKS v1.30 (Catchen *et al.* 2013; Catchen *et al.* 2011). Raw data were demultiplexed and filtered using the "process_rad_tags" script. Average quality scores were determined within a sliding window 15% the length of the sequence, and reads with scores below 90% probability of being correct were discarded. I discarded samples with less than 5,000 polymorphic loci. Barcodes were trimmed resulting in 96 bp sequences, and loci were assembled using the STACKS "de_novo_map.pl" pipeline, using a minimum of three identical reads to create a stack (m= 3), three mismatches allowed between loci within an individual (M= 3), five mismatches when aligning reads (N= 5), and two mismatches when building the catalog (n= 2).

I used the "populations" script to obtain desired subsets of loci. In order to do a hierarchical analysis of population genetic structure, I produced two datasets: 1) a large-scale dataset that included all the sampled populations from the Red Sea and Arabian Peninsula, and 2) a small-scale dataset including only the Red Sea populations. I did that to obtain a powerful subset of loci to independently analyze population structure within the Red Sea, and to make sure the loci shared were polymorphic within the Red Sea (if I manually remove populations of the larger dataset instead of starting with a new filter designed for the Red Sea, some loci may be fixed for the Red Sea but polymorphic in the large scale dataset).

The settings used in the large-scale dataset (Red Sea and Arabian Peninsula) were loci shared between all populations (p= 8), in at least 75% of individuals within a population (r= 0.75) and with coverage of 8x (m= 8). I used the first SNP of each sequence, and removed loci with minor allele frequencies lower than 5%. The filtering resulted in a total of 149 individuals with 2,977 loci and a data matrix 90% complete (see Fig.1.1 for samples per site).

The settings used in the small-scale dataset (Red Sea) were loci shared between all populations (p= 5), in at least 75% of individuals within a population (r= 0.75) and with coverage of 8x (m= 8). I used the first SNP of each sequence, and removed loci with minor allele frequencies lower than 5%. The filtering resulted in a total of 108 individuals with 3,816 loci and a data matrix 89% complete

For all downstream analyses, I used the STRUCTURE output files produced by STACKS, which were converted to other file formats using PGDSPIDER 2.0 (Lischer & Excoffier 2012).

RADSeq data analysis

For the small scale and the large-scale dataset, I did the following: I identified outlier loci using the modified FDIST approach (Excoffier *et al.* 2009) implemented in ARLEQUIN (Excoffier & Lischer 2010). Using these results I classified each locus into one of three categories: 1) divergent outlier loci that have F_{ST} values significantly higher than expected (p-value <0.01) 2) candidate loci under balancing selection that have F_{ST} values significantly lower than expected (p-value <0.01), and 3) neutral loci that include the remaining ones. Population genetic analyses were applied to all loci, neutral loci and divergent loci separately. Balancing loci were not used in data analysis. Hereafter, divergent outlier loci are referred to as “outliers”.

To analyze hierarchical population structure, I applied an Analisis of Molecular Variance (AMOVA) with ARLEQUIN 3.5.1.2 (Excoffier & Lischer 2010). Global and pairwise F_{ST} values were estimated with ARLEQUIN using 10,000 permutations. Genetic assignment of individuals was calculated with

STRUCTURE (Pritchard *et al.* 2000) using correlated allele frequencies in an admixture model, one million Markov chain Monte Carlo (MCMC) repetitions and 100,000 burn-in runs. I ran 10 simulations for each K (from 1 to 9 for the large-scale dataset, from 1-6 for the small-scale dataset). The most likely number of clusters (K) was determined with the Evanno method (Evanno *et al.* 2005) using STRUCTURE HARVESTER (Earl & vonHoldt 2012). DISTRUCT (Rosenberg 2004) was used to generate the graphical display of population structure using the output of CLUMPP (Jakobsson & Rosenberg 2007). A non-Bayesian clustering method was also applied, the discriminant analysis of principal components (DAPC) (Jombart *et al.* 2010), available in ADEGENET (Jombart 2008) for R (R Development Core Team 2015). This multivariate method does not rely on population genetics models and generates a graphical representation of population relationships. The method seeks to show differences between groups as best as possible while minimizing variation within populations (Jombart *et al.* 2010). The best number of principal components (PCs) was identified using the cross validation method ("xValDapc" function in ADEGENET). The "find.clusters algorithm" was also calculated, to investigate the putative number of genetic clusters in the data.

RESULTS

Analysis of loci outliers

In the large-scale dataset (Red Sea and Arabian Peninsula) I obtained a total of 2,977 loci. The analysis identified 68 outliers for positive selection, 43 outliers for balancing selection, and the remaining loci (2,866) were classified as neutral. In the small-scale dataset (Red Sea) I obtained a total of 3,816 loci. The analysis identified 77 outlier loci under positive selection, 35 under balancing selection, and the remaining loci (3,704) were classified as neutral.

Population genomics in the Red Sea and Arabian Peninsula (Large-scale dataset)

In the large-scale dataset AMOVA, the groups defined were the Red Sea (AQA, NRS, CRS, SRS, FAR) and the Arabian Peninsula (DJI, SOC, OMA). AMOVA analysis only revealed significant differences between the Red Sea and Arabian Peninsula when analyzing all the loci (n=2977, Table 1.1) and the outliers (n=68, Table 1.1). Pairwise F_{ST} comparisons showed that Oman was significantly different from the other populations for neutral loci, all loci, and the outliers (Tables 1.2,1.3,1.4). Socotra was different in analysis with all loci and outliers (Tables 1.3,1.4). Northern Red Sea was different in neutral and all loci, but not outliers (Tables 1.2,1.3). Dijbouti was only different in the outlier pairwise F_{ST} comparisons (Table 1.4).

Bayesian population clustering with STRUCTURE showed no differences (k=1) when using all loci or neutral loci (Fig. 1.2). In contrast, analysis with the outliers revealed strong genetic structure. The Evanno method supported the presence of 2 genetic groups: Socotra and all the other populations (for k=2, ΔK=7007). The Evanno method also supported 3 genetic groups: Socotra, Red Sea and a group of Djibouti and Oman (for k=3, ΔK=235). Upon visual inspection of the individual assignments for the two most likely answers, here I show the result with the three genetic groups (Fig. 1.2), because it is informative about the population structure, and in these type of situations when there is clear genetic structure is better not to strictly adhere to the Evanno method or other statistical methods to define the number of genetic clusters (Meirmans 2015).

The DAPC, a method that does not rely on population genetic models, showed similar results to those revealed by STRUCTURE (Fig. 1.3). All the populations formed one group in the neutral loci analysis, and the number of clusters determined by DAPC was one. With all the loci, the DAPC showed that Socotra is the most differentiated of all (Fig. 1.3), followed by Oman and Northern Red Sea, however such differences are not large because the number of clusters determined by DAPC was one. Outlier loci DAPC showed that Socotra is the most differentiated, followed by Oman and Djibouti (Fig. 1.3), and the analysis determined the presence of four genetic clusters.

Population genomics within the Red Sea (Small-scale dataset)

In the small-scale dataset AMOVA, the groups defined were the high-latitude populations (AQA, NRS, CRS) and low-latitude ones (SRS, FAR). AMOVA analysis showed no significant differences between the high latitude and low latitude Red Sea groups that I defined (Table 1.5). This is true for all the analysis (all loci, neutral or positive). Pairwise F_{ST} comparisons for neutral and all loci showed that the Northern Red Sea is significantly different from Aqaba and Central Red Sea (Tables 1.6,1.7). All comparisons were significant with the outliers (n= 77), in this case Southern Red Sea and Farasan had the largest differences with all the other populations (Table 1.8).

Bayesian population clustering with STRUCTURE showed no differences (k= 1) when using all loci or neutral loci (Fig. 1.2). Analysis with the outliers revealed the presence of 3 clusters, supported by the Evanno method (ΔK=235): some individuals from Aqaba and Northern Red Sea are a distinct genetic group (Fig. 1.2). There is a central cluster formed by the majority of Northern Red Sea and Central Red Sea individuals, with some of the Southern Red Sea individuals. There is a southern cluster formed by the majority of Southern Red Sea individuals and Farasan (Fig. 1.2).

The DAPC, a method that does not rely on population genetic models, showed similar results to those revealed by pairwise F_{ST} and STRUCTURE. DAPC with neutral and all loci showed that Aqaba is the most differentiated of all, followed by the Northern Red Sea (Fig. 1.3), however such differences are not large because only one cluster was determined by DAPC. Outlier loci mirror the results found with STRUCTURE, where there are three groups: Aqaba, the Northern Red Sea/ Central Red Sea and the Southern Red Sea/Farasan (Fig. 1.3). Furthermore, the analysis validated the presence of three genetic clusters.

TABLES

Table 1.1. AMOVAs between the Red Sea and Arabian Peninsula

	% Variation	FCT	p-value
Neutral (2,866)	-0.04	-0.0004	0.7321+-0.0046
All (2,977)	10	0.0010	0.0372+-0.018
Outliers (68)	5.93	0.0593	0.0172+-0.0013

Table 1.2. Pairwise F_{ST} in the Arabian Peninsula populations, neutral loci (n=2,866)

	AQA	NRS	CRS	SRS	FAR	DJI	SOC	OMA
AQA	***							
NRS	**0.0046**	***						
CRS	0.0001	**0.0024**	***					
SRS	-0.0010	0.0009	-0.0030	***				
FAR	-0.0023	0.0007	-0.0020	-0.001	***			
DJI	0.0018	**0.0032**	-0.0013	-0.001	-0.001	***		
SOC	0.0005	**0.0028**	-0.0015	-0.002	0.000	**0.0030**	***	
OMA	0.0014	**0.0052**	**0.0015**	**0.002**	**0.004**	**0.0029**	**0.0043**	***

Table 1.3. Pairwise F_{ST} in the Arabian Peninsula populations, all loci (n=2,977)

	AQA	NRS	CRS	SRS	FAR	DJI	SOC	OMA
AQA	***							
NRS	**0.0057**	***						
CRS	0.0007	**0.0032**	***					
SRS	-0.0002	**0.0017**	-0.0019	***				
FAR	-0.0010	**0.0022**	-0.0007	-0.0006	***			
DJI	**0.0037**	**0.0059**	0.0008	0.0007	0.0007	***		
SOC	**0.0056**	**0.0092**	**0.0053**	**0.0039**	**0.0055**	**0.0092**	***	
OMA	**0.0041**	**0.0083**	**0.0044**	**0.0052**	**0.0066**	**0.0048**	**0.0099**	***

Table 1.4. Pairwise F_{ST} in the Arabian Peninsula populations, outliers (n=68)

	AQA	NRS	CRS	SRS	FAR	DJI	SOC	OMA
AQA	***							
NRS	**0.0409**	***						
CRS	**0.0255**	**0.0349**	***					
SRS	**0.0324**	**0.0358**	**0.0518**	***				
FAR	**0.0536**	**0.0641**	**0.0628**	**0.0341**	***			
DJI	**0.0870**	**0.1139**	**0.1040**	**0.0972**	**0.0814**	***		
SOC	**0.1735**	**0.2075**	**0.2360**	**0.1986**	**0.2122**	**0.2168**	***	
OMA	**0.1085**	**0.1285**	**0.1346**	**0.1322**	**0.1332**	**0.0981**	**0.1963**	***

Table 1.5. AMOVAs between the Red Sea regions. High latitude region: AQA, NRS, CRS. Low latitude region: SRS, FAR.

	% Variation	FCT	p-value
Neutral (3,704)	-0.34	-0.0034	1.0000+-0000
All (3,816)	-0.22	-0.0022	1.0000+-0000
Outliers (77)	5.67	0.0567	0.0955+-0000

Table 1.6. Pairwise F_{ST} in the Red Sea populations, neutral loci (n=3,704)

	AQA	NRS	CRS	SRS	FAR
AQA	***				
NRS	**0.0042**	***			
CRS	0.0003	**0.0017**	***		
SRS	-0.0018	-0.0010	-0.0031	***	
FAR	-0.0035	-0.0008	-0.0022	-0.0010	***

Table 1.7. Pairwise F_{ST} in the Red Sea populations, all loci (n=3,816)

	AQA	NRS	CRS	SRS	FAR
AQA	***				
NRS	**0.0051**	***			
CRS	**0.0016**	**0.0025**	***		
SRS	0.0008	0.0006	-0.0018	***	
FAR	-0.0005	0.0013	-0.0001	0.0002	***

Table 1.8. Pairwise F_{ST} in the Red Sea populations, outliers (n=77)

	AQA	NRS	CRS	SRS	FAR
AQA	***				
NRS	**0.0414**	***			
CRS	**0.0639**	**0.0419**	***		
SRS	**0.1148**	**0.0835**	**0.0701**	***	
FAR	**0.1330**	**0.1113**	**0.1069**	**0.0576**	***

FIGURES

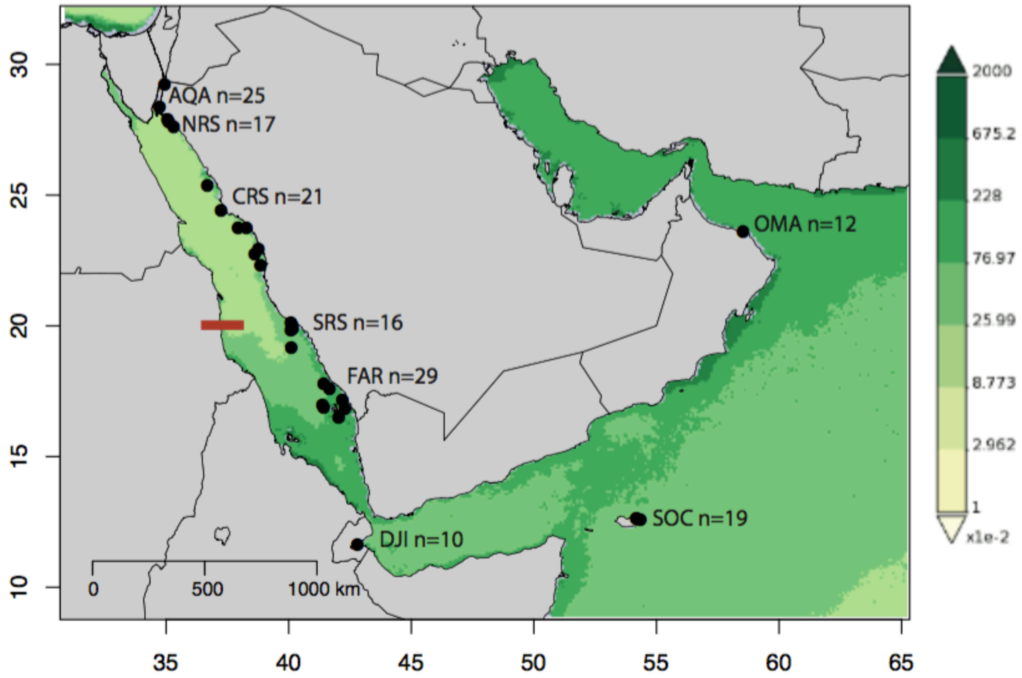

Figure 1.1. Sampling locations and sample sizes. AQA=Aqaba, NRS=Northern Red Sea, CRS=Central Red Sea, SRS=Southern Red Sea, FAR=Farasan islands, DJI=Djibouti, SOC=Socotra, OMA=Oman. The Red line indicates the 20°N of latitude where there is a change in productivity, as shown by Chlorophyll a concentrations in the surface (scale in mg/m^3). There are also eddies and changes in salinity and sea surface temperature near this region. The large-scale dataset included all the populations. The small-scale dataset included AQA, NRS, CRS, SRS, FAR.

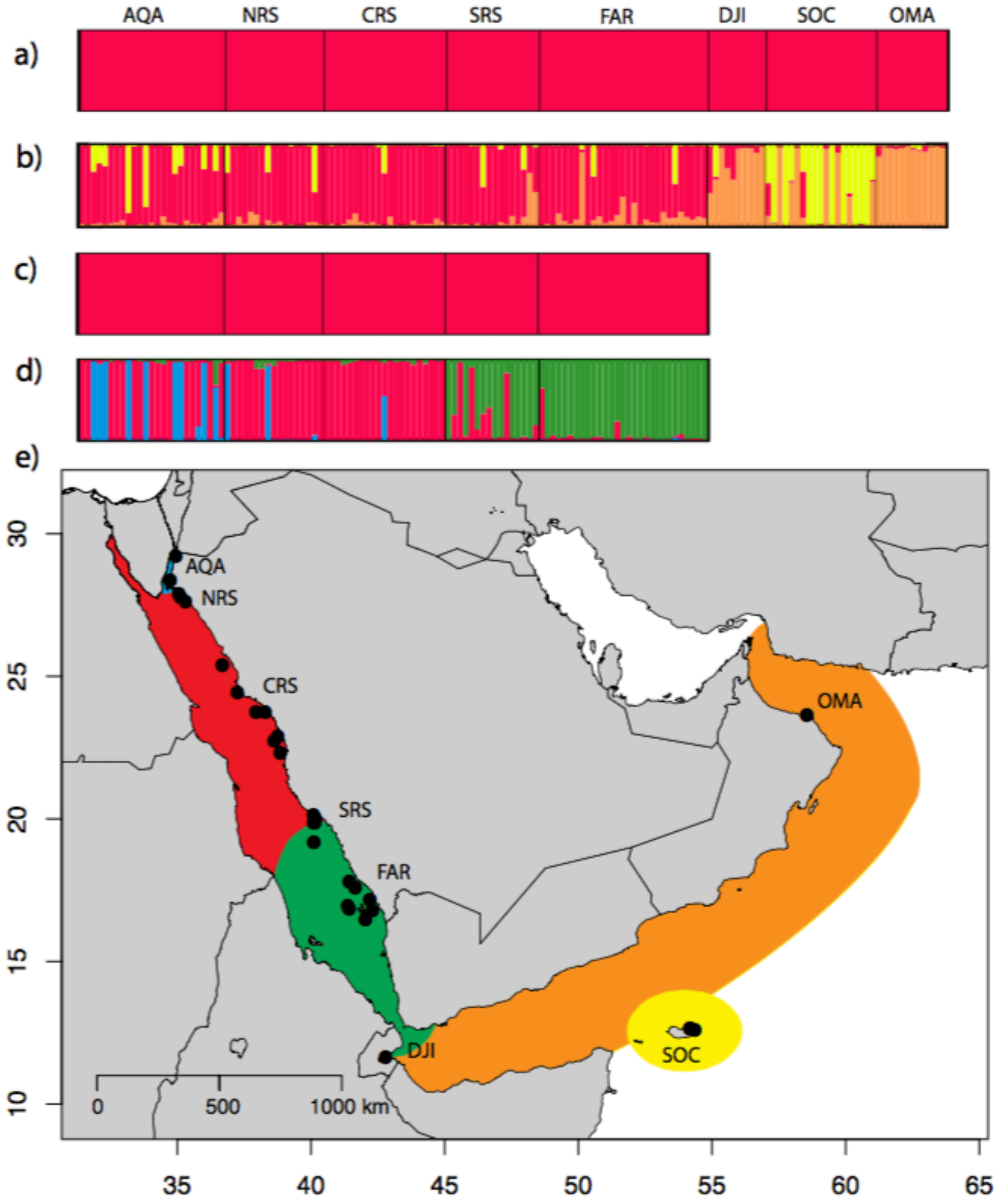

Figure 1.2. Hierarchical Bayesian posterior probability assignment of sampled *Dascyllus trimaculatus*, according to analysis with all loci, outlier loci and neutral loci. (a) Initial analysis using all individuals (n=149) recovered one genetic cluster if I used all loci or neutral loci, (b) but using outlier loci I recovered 3 genetic clusters (ΔK=235). (c) Subsequent analysis within the Red Sea (n=108) recovered one genetic cluster if I used all loci or neutral loci, (d) but using outlier loci I recovered 3 clusters (ΔK=979). (e) Altogether the outlier analysis shows up to 5 genetic groups, and the map illustrates my hypothesis of the geographic patterns of genetic structure within the Arabian Peninsula.

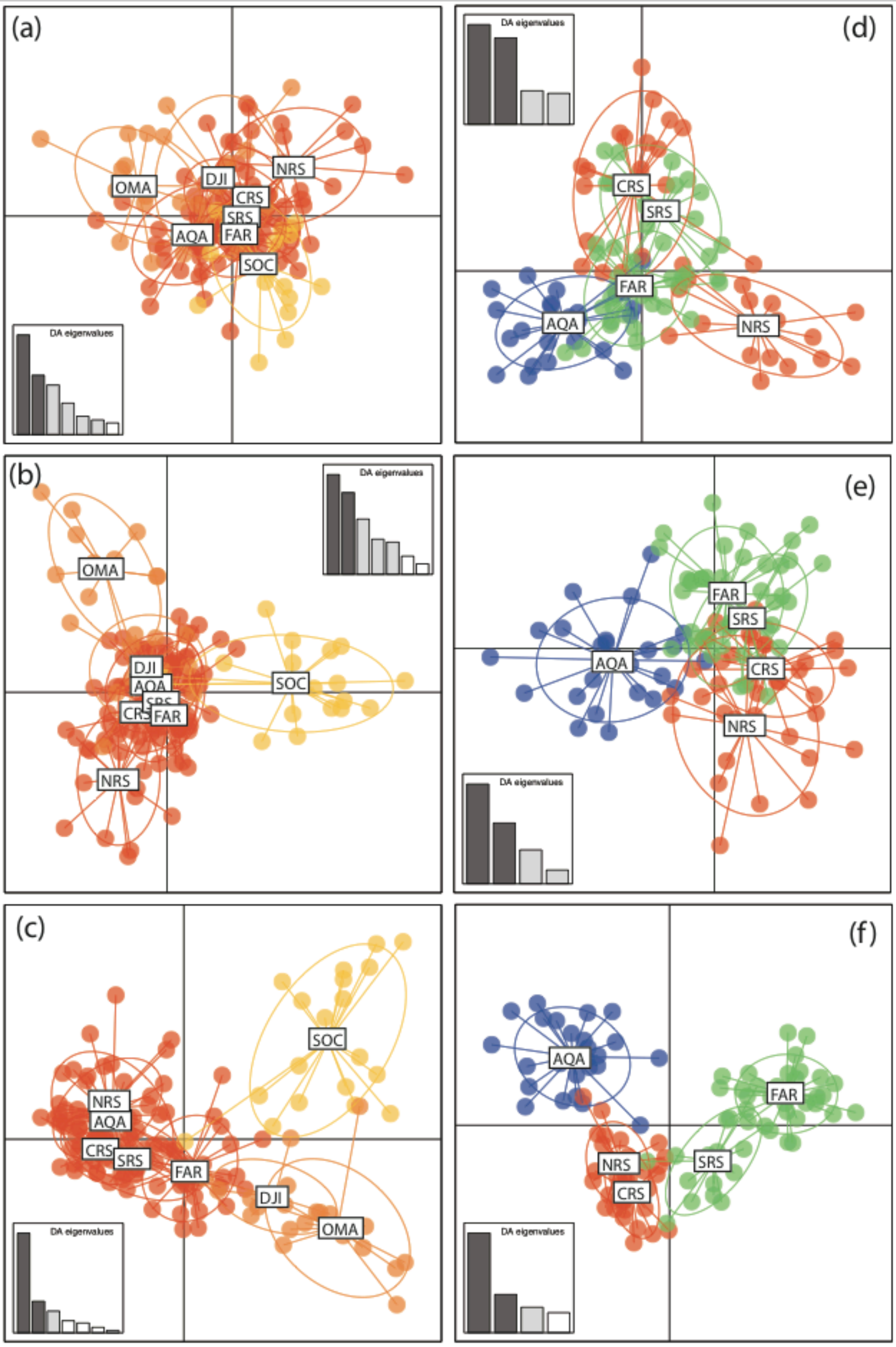

Figure 1.3. Left: Arabian Peninsula DAPC of neutral loci (top), all loci (center) and outliers (bottom). Right: Red Sea DAPC of neutral loci (top), all loci (center) and outliers (bottom).

DISCUSSION

I found hierarchical genetic structure consistent with expected barriers to connectivity in the region: 1) Within the Arabian Peninsula, I found that Socotra populations were highly differentiated, followed by a less differentiated population cluster outside of the Red Sea conformed by Djibouti and Oman. 2) Within the Red Sea, I found one genetic break near 20° N, and I also found genetic differentiation of Aqaba and Northern Red Sea with other populations; Nevertheless, the genetic structure is only detected in the set of outlier loci.

Contrasting results between outliers and neutral loci have been found in other studies. For example, Gaither *et al.* (2015) in *Acanthurus spp*. surgeonfishes, and Bernardi *et al.* (2016) in the bluespotted cornetifish *Fistularia commersonii*. The authors did similar analysis and found no evidence of population structure with the neutral loci, but they found strong fixation concordant with meaningful patterns such as expected genetic breaks or genetic changes due to population invasions. In their case studies the genetic differences were attributed to selection. In this study, the patterns found could be attributed to selection in the outliers, and to incomplete allele sorting in the neutral loci. *Dascyllus trimaculatus* is a common and widespread species that lives from the Red Sea to French Polynesia. It forms a species complex of very closely related species that diverged from the *Dascyllus reticulatus* complex very recently during the Pleistocene, about 3.9 mya (Bernardi & Crane 1999; McCafferty *et al.* 2002). Its highest genetic diversity is reached in the Indian Ocean (Leray *et al.* 2010). It reproduces year-round, up to three times per month, and the egg clutches are estimated at 20,000-25,000 (Randall & Allen 1977). Its pelagic larval duration in the Red Sea ranges from 20-27 days (Robitzch *et al.* 2016). Thus, due to its life history characteristics, its recent divergence and high effective population sizes, it is not surprising to find low genetic divergence in neutral markers.

Are the patterns found in the outliers of special significance to the species, or is the most important result the low genetic structure that is indicated by the neutral markers? Are these patterns reflected only in outliers because the environmental gradients affect the fitness of the species? These are many of the questions that remain to be answered. Here, I focus the discussion on a comparison with findings of Red Sea connectivity with other species, the discrepancies between studies, and the potential mechanisms explaining the patterns.

The genetic breaks found in the outlier loci between the Arabian Peninsula and the Red Sea are consistent with biogeographic patterns in the region. In terms of coral reef-associated fishes, there are significant differences between the Red Sea, Gulf of Aden, and Arabian Gulf for families such as Chaetodontidae, Pomacanthidae, Pomacentridae, Acanthuridae, Scaridae, Labridae, Lethrinidae and Lutjanidae (Khalaf & Kochzius 2002). Differences in connectivity and community composition between the Red Sea and the Arabian Peninsula have been attributed to two barriers: the shallow and narrow opening of the Red Sea at the Strait of Bab el Mandab, which allows little exchange between them

(Klausewitz 1989; Roberts *et al.* 1992), and the lack of reef development and environmental conditions along the southern coast of Somalia and Oman. During the summer, the Somali Current brings upwelling to southern Oman and Somalia (Kemp 1998), a region with limited abundance of coral reef. My data supports the idea that there is isolation between the Red Sea and the Arabian Peninsula populations, and that Socotra is a highly differentiated location. It is likely that populations in Socotra are more related to the rest of the Western Indian Ocean, as there are strong genetic differences between populations in the Red Sea/Arabian Peninsula and outside (Salas *et al.*, in prep.). In addition, Socotra is considered a marine suture zone, located at the boundaries the Red Sea/Arabian Peninsula and the Western Indian Ocean province (DiBattista *et al.* 2015). The genetic cluster found in Socotra has some individuals that assign to the Arabian Peninsula (Fig 1.2), suggesting that both Western Indian Ocean and Arabian Peninsula faunas meet in this interesting region.

Within the Red Sea, I found three genetic clusters; a group located in Aqaba with some spread in the Northern Red Sea, another group including most of the Northern Red Sea, the Central Red Sea and some individuals from the Southern Red Sea, and a third group including some individuals from the Southern Red Sea and Farasan Islands (Fig. 1.2). This last delineation is a genetic break near 20° N. Differences of fish and coral community structure have been attributed to changes in habitat and an abrupt increase in turbidity south of 20° N (Roberts *et al.* 1992; Sheppard & Sheppard 1991). The reefs sit in a shallow platform and the waters are generally warmer and with higher productivity, as revealed by chlorophyll a concentrations (Fig. 1.1). This environmental change affects the pelagic larval duration (PLD) of the three species of *Dascyllus* found in the Red Sea, where for all species, PLDs follow a latitudinal gradient; PLDs are larger in Aqaba and shorter in the Farasan islands. These differences are correlated with latitude, chlorophyll a concentrations and sea surface temperature (Robitzch *et al.* 2016). As suggested by Robitzch *et al.* (2016) *Dascyllus trimaculatus* larvae may have higher food availability and higher metabolism in this region, causing them to settle from the plankton faster. Adults may have more food and higher metabolism as well, adults feed on plankton: copepods and have been found in their stomach contents (Randall & Allen 1977). Therefore the environmental conditions in the Red Sea could be creating different adaptive regimes all the way from Aqaba to the Farasan. Other studies have found genetic breaks associated with environmental variables (Giles *et al.* 2015; Nanninga *et al.* 2014). It is interesting to note that the southern genetic break detected in this study is not in the same location as that found in these other studies. The break they detected is south of 20° N, located at the Farasan Islands, while mine starts north of the Farasan Islands, near Al-lith. The southern genetic break was identified for the endemic clownfish *Amphiprion bicintus* and for the sessile sponge, *Stylissa carteri*.

Steep environmental gradients, dispersal barriers and patchy habitat distribution can generate phylogeographic breaks in the ocean. Not all species will show genetic breaks. For example, within the Red Sea, connectivity studies found no evidence of a genetic break in the coral, *Pocillopora verrucosa* (Robitzch *et al.*

2015). Of species for which genetic breaks have been detected, they are not always in the exact same location. Genetic breaks often align with recognized biogeographic barriers in each region only on a coarse spatial scale (Pelc *et al.* 2009). Conditions that generate genetic breaks can change over long time scales as barriers shift, but the original signal of historical separation is more likely to persist in species with low levels of gene flow (Hellberg *et al.* 2002). In contrast, the historical signal may be erased or may shift in species with high levels of gene flow (Bertness *et al.* 2001), and dispersal by ocean currents can relocate the positions of genetic breaks (Endler 1977; Pringle & Wares 2007). Due to all these reasons, there may be discrepancies between species with different life histories (Pelc *et al.* 2009). Clownfishes and sponges have lower dispersal distances than *D. trimaculatus* and this may account for the lack of detection of this break in neutral loci, and the different location of the genetic break of the outliers of *D. trimaculatus* when compared with other species.

In conclusion, I found hierarchical genetic structure in populations of *Dascyllus trimaculatus* in the Red Sea and Arabian Peninsula that are reflected in the loci outliers but absent from neutral markers. The Arabian Peninsula and Red Sea may be divergent and relatively isolated due to environmental conditions near the mouth of the Red Sea and upwelling and lack of reef habitat between Somalia and Oman. Within the Red Sea, I found coarse concordance with previously identified genetic breaks but not exactly in the same locations as found for shorter dispersal species. The patterns of genetic differentiation found in the outliers may be due to adaptation to the steep environmental gradient in the Red Sea, and future studies should investigate what mechanisms and what genes are involved in these changes.

REFERENCES

Almany G, Berumen M, Thorrold S, Planes S, Jones G (2007) Local replenishment of coral reef fish populations in a marine reserve. *Science* **316**, 742.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD-genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species; the American lobster (*Homarus americanus*). *Molecular Ecology* **24**, 3299-3315.

Bernardi G, Azzurro E, Golani D, Miller MR (2016) Genomic signatures of rapid adaptive evolution in the bluespotted cornetfish, a Mediterranean Lessepsian invader. *Molecular Ecology*.

Bernardi G, Crane N (1999) Molecular phylogeny of the humbug damselfishes inferred from mtDNA sequences. *Journal of Fish Biology* **54**, 1210-1217.

Bertness MD, Gaines SD, Hay ME (2001) *Marine community ecology* Sinauer Associates Sunderland, Massachusetts.

Botsford LW, Brumbaugh DR, Grimes C, Kellner JB, Largier J, O'Farrell MR, Ralston S, Soulanille E, Wespestad V (2009) Connectivity, sustainability, and yield: bridging the gap between conventional fisheries management and marine protected areas. *Reviews in Fish Biology and Fisheries* **19**, 69-95.

Bowler DE, Benton TG (2005) Causes and consequences of animal dispersal strategies: relating individual behaviour to spatial dynamics. *Biological Reviews* **80**, 205-225.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.

Christie M, Johnson D, Stallings C, Hixon M (2010) Self-recruitment and sweepstakes reproduction amid extensive gene flow in a coral-reef fish. *Molecular Ecology* **19**, 1042-1057.

DiBattista JD, Rocha LA, Hobbs JPA, He S, Priest MA, Sinclair-Taylor TH, Bowen BW, Berumen ML (2015) When biogeographical provinces collide: hybridization of reef fishes at the crossroads of marine biogeographical provinces in the Arabian Sea. *Journal of Biogeography* **42**, 1601-1614.

Dieckmann U, O'Hara B, Weisser W (1999) The evolutionary ecology of dispersal. *Trends in Ecology & Evolution* **14**, 88-90.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Endler JA (1977) *Geographic variation, speciation, and clines* Princeton University Press.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Gadgil M (1971) Dispersal: population consequences and evolution. *Ecology* **52**, 253-261.

Gaither MR, Bernal MA, Coleman RR, Bowen BW, Jones SA, Simison WB, Rocha LA (2015) Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Molecular Ecology* **24**, 1543–1557.

Giles EC, Saenz-Agudelo P, Hussey NE, Ravasi T, Berumen ML (2015) Exploring seascape genetics and kinship in the reef sponge *Stylissa carteri* in the Red Sea. *Ecology and Evolution* **5**, 2487-2502.

Haney R, Silliman B, Rand D (2007) A multi-locus assessment of connectivity and historical demography in the bluehead wrasse (Thalassoma bifasciatum). *Heredity* **98**, 294-302.

Hellberg ME (2009) Gene flow and isolation among populations of marine animals. *Annu. Rev. Ecol. Evol. Syst.* **40**, 291-310.

Hellberg ME, Burton RS, Neigel JE, Palumbi SR (2002) Genetic assessment of connectivity among marine populations. *Bulletin of Marine Science* **70**, 273-290.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

Jones G, Milicich M, Emslie M, Lunow C (1999) Self-recruitment in a coral reef fish population. *Nature* **402**, 802-804.

Jones G, Planes S, Thorrold S (2005) Coral reef fish larvae settle close to home. *Current Biology* **15**, 1314-1318.

Kemp J (1998) Zoogeography of the coral reef fishes of the Socotra Archipelago. *Journal of Biogeography* **25**, 919-933.

Khalaf MA, Kochzius M (2002) Community structure and biogeography of shore fishes in the Gulf of Aqaba, Red Sea. *Helgoland marine research* **55**, 252-284.

Klausewitz W (1989) Evolutionary history and zoogeography of the Red Sea ichthyofauna. *Fauna of Saudi Arabia* **10**, 310-337.

Leray M, Beldade R, Holbrook S, Schmitt R, Planes S, Bernardi G (2010) Allopatric divergence and speciation in a coral reef fish: The three-spot dascyllus, *Dascyllus trimaculatus* species complex. *Evolution* **64**, 1218-1230.

Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

McCafferty S, Bermingham E, Quenouille B, Planes S, Hoelzer G, Asoh K (2002) Historical biogeography and molecular systematics of the Indo-Pacific genus *Dascyllus* (Teleostei: Pomacentridae). *Molecular Ecology* **11**, 1377-1392.

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology* **24**, 3223-3231.

Nanninga GB, Saenz-Agudelo P, Manica A, Berumen ML (2014) Environmental gradients predict the genetic population structure of a coral reef fish in the Red Sea. *Molecular Ecology* **23**, 591–602.

Pechenik JA (1999) On the advantages and disadvantages of larval stages in benthic marine invertebrate life cycles. *Marine Ecology-Progress Series* **177**, 269-297.

Pelc R, Warner R, Gaines S (2009) Geographical patterns of genetic structure in marine species with contrasting life histories. *Journal of Biogeography* **36**, 1881-1890.

Pringle JM, Wares JP (2007) Going against the flow: maintenance of alongshore variation in allele frequency in a coastal ocean. *Marine Ecology Progress Series* **335**, 69-84.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945.

Purcell JF, Cowen RK, Hughes CR, Williams DA (2006) Weak genetic structure indicates strong dispersal limits: a tale of two coral reef fish. *Proceedings of the Royal Society B: Biological Sciences* **273**, 1483-1490.

Randall HA, Allen GR (1977) A revision of the damselfish genus Dascyllus (Pomacentridae) with the description of a new species. *Records of the Australian Museum* **31**, 349-385.

Roberts CM, Shepherd ARD, Ormond RF (1992) Large-scale variation in assemblage structure of Red Sea butterflyfishes and angelfishes. *Journal of Biogeography*, 239-250.

Robitzch V, Banguera-Hinestroza E, Sawall Y, Al-Sofyani A, Voolstra CR (2015) Absence of genetic differentiation in the coral Pocillopora verrucosa along environmental gradients of the Saudi Arabian Red Sea. *Frontiers in Marine Science* **2**, 5.

Robitzch V, Lozano-Cortes D, Kandler N, Salas E, Berumen ML (2016) Productivity and sea surface temperature are correlated with the pelagic larval duration of damselfishes in the Red Sea. *Marine Pollution Bulletin* **105**, 566-574.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Saenz-Agudelo P, Dibattista JD, Piatek MJ, Gaither MR, Harrison HB, Nanninga GB, Berumen ML (2015) Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Molecular Ecology* **24**, 6241-6255.

Sheppard CR, Sheppard AS (1991) *Corals and coral communities of Arabia*. vol 12. Fauna of Saudi Arabia. 170pp.

Swearer S, Caselle J, Lea D, Warner R (1999) Larval retention and recruitment in an island population of a coral-reef fish. *Nature* **402**, 799-802.

Weersing K, Toonen RJ (2009) Population genetics, larval dispersal, and connectivity in marine systems. *Marine Ecology Progress Series* **393**, 1-12.

Chapter 2

Genomics reveals regional population structure for a coral reef fish

CHAPTER 2

GENOMICS REVEALS REGIONAL POPULATION STRUCTURE OF A CORAL REEF FISH

ABSTRACT

Population genetic tools are commonly used to infer connectivity between distant populations. Here, I investigate the population genomic structure of the reef fish *Dascyllus trimaculatus* in the Red Sea, Arabian Sea and Western Indian Ocean, using SNPs generated with RADseq. Neutral loci revealed a signature of gene flow between distant populations, with only subtle and weakly supported genetic differentiation between the Northwestern (Red Sea and Arabian Sea) and Western Indian Ocean provinces. Outlier loci reveal a similar pattern but with a much stronger distinction between provinces that was detected across multiple genetic analysis. In this case, outlier loci clearly resolve the genetic division between provinces and prove to be useful in the detection of subtle genetic structure. While I could not clearly identify the mechanism (isolation, adaption, or both) driving these patterns the data are important to the conservation and management of reef organisms in the region as they indicate that population level divergences are largely concordant with provincial distinctions based on species composition. This dataset also highlights the utility of outlier loci in studies of population connectivity, where large effective population sizes can blur the signals of genetic structure, as is common in a diversity of taxa residing on coral reefs.

INTRODUCTION

Gene flow between geographically disconnected populations is a fundamental process driving the evolution, distribution, and maintenance of species diversity (Bowler & Benton 2005; Kinlan & Gaines 2003). Coral reef fishes generally disperse during a pelagic larval stage (Leis 1991), potentially traveling hundreds of kilometers away from their natal reef (Victor 1991). Understanding how populations are connected by dispersal is vital to conservation efforts, however, it is logically unfeasible to tag and recapture microscopic larvae, or to carry out the extensive sampling needed for parentage studies over large spatial scales (>1,000 km). Instead, population genetics is commonly used to investigate connectivity patterns between distant populations (Lowe & Allendorf 2010).

Marine taxa tend to have large effective population sizes (Ward *et al.* 1994) with only a few migrants per generation sufficient to counter genetic divergence (Palumbi 2003). While low genetic differentiation may have important evolutionary consequences, it may have less of an effect on population demographics that is important to fisheries management and the design of marine protected areas (Benestan *et al.* 2015; Palumbi 2003). Traditional connectivity studies employed a targeted locus approach (e.g., mitochondrial DNA, microsatellites, a few introns), and often showed weak genetic structure and limited resolution (Waples *et al.* 2008). However, with the advent of high throughput sequencing scientists can sample larger numbers of loci which offers increased resolution to detect low levels of genetic structure (Rocha 2013) and in some cases identify regions of the genome under selection that may be linked to local adaptation (Gaither *et al.* 2015).

Here I investigated the population genomics of a coral reef fish in the Red Sea, Arabian Sea and Western Indian Ocean. As per Kulbicki *et al.*, (2013), this region is divided into the Northwestern Indian Ocean Province (NWIOP) and the Western Indian Ocean Province (WIOP) (Fig 2.1). Whereas most studies in the region have focused on comparisons between the Indian and Pacific Oceans (Gaither & Rocha 2013; Ridgway & Sampayo 2007), studies on population connectivity within the Indian Ocean are just starting to flourish e.g. (Ahti *et al.* 2016; DiBattista *et al.* 2013; Fernandez-Silva *et al.* 2015; Visram *et al.* 2010)

My study species, the three-spot dascyllus, *Dascyllus trimaculatus,* is an abundant coral reef fish found throughout the Indo-Pacific (Bernardi *et al.* 2001). It has a pelagic larval stage that lasts an average of 26 days (Wellington & Victor 1989). Juveniles live mostly on anemones while adults tend to stay close to, but are not strictly associated with the anemones (Bernardi *et al.* 2001). Previous studies, based on the mitochondrial control region and microsatellites, showed a single well mixed and genetically diverse Indian Ocean population that was genetically distinct from populations in the Pacific Ocean (Bernardi *et al.* 2002; Leray *et al.* 2010). Here I used single nucleotide polymorphisms (SNPs) developed using restriction-site associated DNA sequencing (RADseq) to assess population structure in Indian Ocean *D. trimaculatus* and to consider the relative contribution of neutral and adaptive processes in partitioning diversity in this geographic region. I hypothesize that Indian Ocean *D. trimaculatus* are not a single panmictic population, and that increased resolution will reveal genetic differentiation across the region.

MATERIALS AND METHODS

Sample collection

A total of 93 individuals from seven populations were collected with pole spears and hand nets by scuba divers, between 1998 and 2013 (Fig 2.1). Populations were grouped into either the NWIOP (Northern Red Sea: NRS, Djibouti: DJI and Oman: OMA) or the WIOP (Diego de Garcia in the Chagos Archipelago: DGA, Zanzibar: ZAN, Mayotte: MAY and Juan de Nova located in the Scattered Islands: JNO, Fig 2.1) according to Kulbicki *et al.*, (2013). The NRS population consisted of 6 individuals from Eilat and 3 from Jeddah, Saudi Arabia. Some of the samples were part of earlier studies (Bernardi *et al.* 2002; Leray *et al.* 2010), including individuals from Eilat, Oman, Mayotte, and Zanzibar.

RADseq library preparation and sequencing

Genomic DNA was extracted using the Qiagen DNeasy animal blood and tissue kit (Qiagen, Valencia, USA). The library was prepared using the double-digest RADseq protocol (Peterson *et al.* 2012), with modifications (see supplementary materials) and sequenced on a single Illumina HiSeq 2000 lane, at the UCLA Neuroscience Genomics Core facility. Raw data was de-multiplexed, quality filtered and trimmed to 95 bp, using the "process_rad_tags" script available in STACKS v1.09 (Catchen *et al.* 2013). Loci were assembled using the STACKS "de novo_map.pl" pipeline while the "populations" script was used to filter loci and create output files. The quality control and filtering resulted in a total of 1,174 loci and a data matrix that was 84% complete (supplemental materials). I used PGDSPIDER 2.0 (Lischer & Excoffier 2012) to convert the resulting STRUCTURE files into other formats.

Data analysis

To identify outlier loci I used the modified FDIST approach (Excoffier *et al.* 2009) implemented in ARLEQUIN (EXCOFFIER & LISCHER 2010). Using these results I classified each locus into one of three categories: 1) divergent outlier loci which have F_{ST} values significantly higher than expected (p-value <0.01) 2) loci under balancing selection which have F_{ST} values significantly lower than expected (p-value <0.01), and 3) neutral loci which include all other loci.

To test for genetic structure I conducted hierarchical AMOVAs and calculated pairwise F_{ST} values (Weir & Cockerham 1984) using ARLEQUIN. Discriminant analyses of principal components (DAPC) (Jombart *et al.* 2010) were executed using ADEGENET (Jombart 2008) for R (R Development Core Team 2015). In addition, I ran the Bayesian clustering method implemented in STRUCTURE. For the latter, the most likely number of clusters (K) was determined using the Evanno method (Evanno *et al.* 2005) of STRUCTURE HARVESTER (Earl & vonHoldt 2012). To test for Isolation by distance (IBD) I compared matrixes of $F_{\mathrm{ST}}/(1\text{-}F_{\mathrm{ST}})$ and minimum ocean distance with Mantel tests performed using GENEPOP (Raymond & Rousset 1995). For details see supplementary materials.

RESULTS

A total of 1,174 loci were obtained for 93 individuals. The FDIST method identified 25 outlier loci and 32 loci putatively under balancing selection. Based on these results I generated three datasets for analysis: 1) neutral loci (n=1,117); 2) outlier loci (n=25); 3) all loci (n=1,174). Results of the dataset with all loci are shown in the supplementary materials.

Based on the neutral loci I found high genetic connectivity and a weak differentiation between the two provinces. Global F_{ST} was low but significant (0.0057, p<0.0001), and 12 of the 21 pairwise F_{ST} comparisons were significant. In particular the population in Oman (OMA) stood out and was significantly different from all other populations (Table 2.1). The AMOVA showed a low but significant divergence between the NWIOP and the WIOP (F_{CT} =0.0041, p=0.0283). The DAPC analysis partitioned the data into only one cluster, and the analysis shows close relationships among all populations except OMA (Fig 2.1). STRUCTURE analysis indicated K=1 (Fig 2.1), which is in agreement with the DAPC results. Finally there was an indication of IBD but the relationship was not significant (Fig 2.1).

Analyzing the outliers (n=25) I found strong genetic differentiation between provinces. Global F_{ST} for the outlier loci was 0.3271 ($p<0.0001$), and all of the pairwise comparisons were significant except between ZAN and JNO (Table 2.1). An AMOVA supports the distinction between the NWIOP and WIOP provinces (F_{CT}=0.2349, p=0.0342). The DAPC analysis identified 16 clusters and separation of the NWIOP and WIOP provinces (Fig 2.1). Analysis of the STRUCTURE results suggests the presence of two clusters (K=2) that closely match the DAPC results (Fig 2.1). Finally, outliers revealed significant IBD (Fig 2.1).

DISCUSSION

Using the increased resolution of 1,174 SNPs I detected significant population structure between Northwestern and Western Indian Ocean *D. trimaculatus,* showing that they are not a single panmictic population. The neutral loci indicated little population structure across the Indian Ocean, with weak genetic differentiation between the provinces. However, using the outlier loci I detected a substantially stronger signal between provinces that was consistent across analyses. These results indicate that the degree of gene flow (i.e., the transfer of alleles from one population to another) is variable across the genome, and may be affected by different processes and/or operate at different scales.

Due to the high dispersal potential and large effective population sizes of most marine fishes, population connectivity can be maintained over large spatial scales and only a few successful colonizing events are needed to greatly reduce genetic differentiation (Kinlan & Gaines 2003; Palumbi 2003). *D. trimaculatus* has an average pelagic larval duration (PLD) of 26 days (Wellington & Victor 1989), which could be sufficient to maintain low genetic differentiation with infrequent long distance or stepping stone dispersal. These few colonizers may be important to populations on evolutionary time scales, but are not necessarily sufficient to allow population persistence at ecological timescales (Cowen & Sponaugle 2009). The neutral dataset show signals of restricted gene flow indicating isolation between provinces, but the signature is hard to detect due to episodic dispersal across provinces. Among the outlier loci the signal is much stronger. Therefore, a signature of isolation is detected in the neutral loci and the outlier results could be driven by a number of mechanisms including isolation, selection or both.

Habitat discontinuities, deep-water upwellings, and the direction of prevailing currents (Fig 2.1), could be responsible for contemporary isolation between the Northwestern and the Western Indian Ocean (Fig 2.1). Seasonal upwellings bring cold and nutrient-rich waters to southern Oman and the Somalian Coast, creating large areas unsuitable for the development of coral reef habitat (Kemp 1998), while currents and complex topography may divert larvae and prevent dispersal between provinces (Fig 2.1). If the divergences revealed here in the outlier loci are due to isolation and not adaptation, then these loci should be subject to the effects of drift and show similar patterns to the neutral loci. In fact the outlier dataset does show similar but stronger signals compared to the neutral loci. AMOVAs based on the neutral and outlier loci demonstrate weak but significant structure between provinces while IBD was found in both the neutral and outlier datasets (albeit not significant in the neutral loci).

Genetic isolation can also have an historical layer to it. The Northwestern Indian Ocean is on the periphery of the Indo-Pacific biogeographic region (DiBattista *et al.* 2013). During Pleistocene glacial cycles the Red Sea and Persian Gulf were subject to periods of extreme isolation when the sea level dropped as much as 130 m below current sea level (DiBattista *et al.* 2016a). In some cases isolation lead to speciation in peripheral populations, while in others it only led to population differentiation, as seen here in *D. trimaculatus*. After the last glacial maxima 26.5 to 19 ka (Clark *et al.* 2009) most populations began to expand as habitat opened up. When a subset of individuals at the frontier of a population expansion move into new territory, their particular alleles increase in frequency, like a wave of genetic drift happening at the expanding population front. Such phenomenon is called “allele surfing” (Excoffier & Ray 2008). Unlike most other demographic effects, surfing generally does not affect all loci, so it can impact neutral allele frequencies in ways that mimic the patterns of directional selection (Excoffier & Ray 2008; Kirk & Freeland 2011) and could be responsible for the results that are more evident in outlier loci.

The Northwestern Indian Ocean is one of the most variable and environmentally extreme regions in the tropical oceans (DiBattista *et al.* 2016a), in contrast with the more stable WIOP ; such differences could be selecting for different traits in *D. trimaculatus* and other species across provinces. During the summer months, the waters between Iran and the Arabian Peninsula become the world's hottest sea, while in the winter they become one of the coldest environments for coral reef growth (Riegl & Purkis 2012). The Red Sea experiences large spatio-temporal fluctuations in physical conditions, and a unique north-south environmental gradient in salinity, temperature and primary productivity (DiBattista *et al.* 2016a). Reefs in both the Red Sea and Oman are known to have high variability in environmental factors such as temperature and salinity (Cavalcante *et al.* 2016; DiBattista *et al.* 2016b; Thoppil & Hogan 2009). Adaptation to these highly variable environments is thought to drive the high rates of endemism in the region (DiBattista *et al.* 2016a) and may affect successful establishment of larvae from non-native populations. There is a possibility that the loci outliers are under selection and reflecting adaptive divergence, however I could not identify responsible genes nor exclude the possibility of false positives (Table 2.S2).

The concordance in patterns between the neutral loci and outliers suggest that the main mechanism is isolation. While it is difficult to distinguish between divergence driven by selection versus divergence driven by isolation, it is important to note that these processes are not mutually exclusive and in fact could be acting in concert on populations found around the region with its complex geologic history and heterogeneous environment. It is possible that physical barriers between the provinces are semipermeable, allowing for restricted dispersal, and environmental contrasts between provinces reinforce those barriers through selection. Historical isolation could also have a role promoting adaptation. Even though it has only been shown in asexual microbial strains, population expansions after isolation can promote adaptation if colonizing individuals carry beneficial mutations (Gralka *et al.* 2016). In my view, all these processes can be happening in the study region and the specific roles of selection and isolation will only be disentangled with the use of experiments.

This study highlights the power of RADSeq loci to reveal cryptic patterns of genetic structure, but the mechanisms driving these patterns are still difficult to resolve. Isolation of the provinces is definitively happening and perhaps in conjunction with adaptive processes. Large effective population sizes in marine taxa have always made it difficult to interpret genetic structure, but outlier loci can help delineate a signal. These results confirm that the Indian Ocean *Dascyllus trimaculatus* are not a single panmictic population but instead there are distinctions between the Northwestern and Western Indian Ocean populations. Despite the lack of a clear mechanism, data like these are important to the conservation and management of reef organisms in the region (Funk *et al.* 2012). These results indicate that the Red Sea and Arabian stocks should be managed separately from the greater Western Indian Ocean stock, and the role of adaptive vs. neutral variation must be examined further.

TABLES

Table 2.1. F_{ST} values between WIO populations, for neutral loci (below asterisks) and outlier divergent loci (above asterisks). Significant values ($p<0.05$) are bolded.

	NRS	DJI	OMA	DGA	ZAN	MAY	JNO
NRS	***	**0.1522**	**0.1919**	**0.2944**	**0.3139**	**0.3030**	**0.3080**
DJI	**0.0055**	***	**0.1825**	**0.2889**	**0.3292**	**0.3516**	**0.3533**
OMA	**0.0046**	**0.0075**	***	**0.3469**	**0.3399**	**0.3768**	**0.3420**
DGA	**0.0056**	**0.0088**	**0.0157**	***	**0.0738**	**0.0714**	**0.1309**
ZAN	0.0030	0.0016	**0.0060**	-0.0016	***	**0.0870**	0.0273
MAY	-0.0033	**0.0050**	**0.0063**	0.0006	-0.0022	***	**0.0821**
JNO	0.0008	**0.0067**	**0.0039**	**0.0018**	-0.0010	-0.0038	***

FIGURES

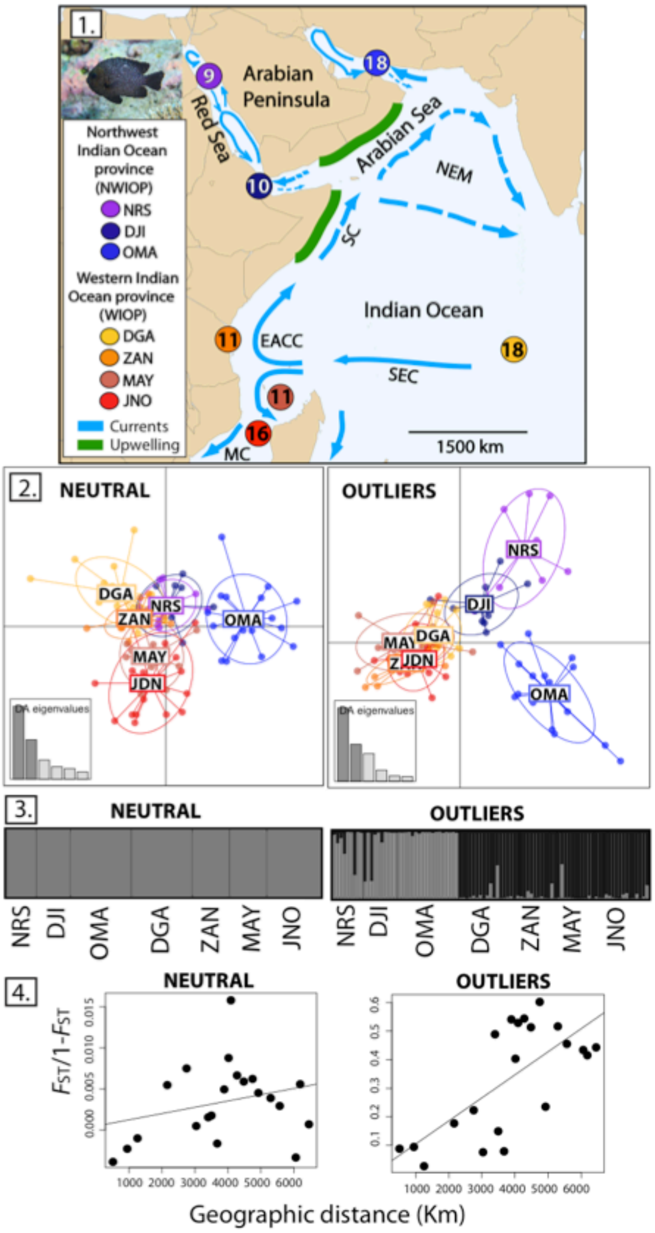

Figure 2.1. Panel 1: Populations and sample sizes (inside circles). NRS: Northern Red Sea; DJI: Djibouti; OMA: Oman; DGA: Diego de García, Chagos; ZAN: Zanzibar; MAY: Mayotte; JDN: Juan de Nova, Scattered Islands. Summer upwelling and currents are shown; dashed lines indicate winter reversals. Currents (C.): NEM: NE Monsoon C., SC: Somali C., EACC: East African Coastal C., MC: Mozambique C., SEC: South Equatorial C. Panel 2: DAPC for neutral loci n= 1,117, and outlier divergent loci n=25. Panel 3: STRUCTURE plot with most likely k for neutral (k=1) and outliers (k=2). Panel 4: Isolation by distance (IBD) in neutral markers, mantel test p=0.1990, r^2=0.0756, and outliers, p=0.016, r^2=0.4801. *D. trimaculatus* picture by Tane Synclair-Taylor.

REFERENCES

Ahti PA, Coleman RR, DiBattista JD, Berumen ML, Rocha LA, Bowen BW (2016) Phylogeography of Indo-Pacific reef fishes: sister wrasses Coris gaimard and C. cuvieri in the Red Sea, Indian Ocean and Pacific Ocean. *Journal of Biogeography* **43**, 1103-1115.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD-genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species; the American lobster (*Homarus americanus*). *Molecular Ecology* **24**, 3299-3315.

Bernardi G, Holbrook SJ, Schmitt RJ (2001) Gene flow at three spatial scales in a coral reef fish, the three-spot dascyllus, *Dascyllus trimaculatus*. *Marine Biology* **138**, 457-465.

Bernardi G, Holbrook SJ, Schmitt RJ, Crane NL, DeMartini E (2002) Species boundaries, populations and colour morphs in the coral reef three–spot damselfish (*Dascyllus trimaculatus*) species complex. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **269**, 599-605.

Bowler DE, Benton TG (2005) Causes and consequences of animal dispersal strategies: relating individual behaviour to spatial dynamics. *Biological Reviews* **80**, 205-225.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Cavalcante GH, Feary DA, Burt JA (2016) The influence of extreme winds on coastal oceanography and its implications for coral population connectivity in the southern Arabian Gulf. *Marine Pollution Bulletin* **105**, 489-497.

Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM (2009) The last glacial maximum. *Science* **325**, 710-714.

Cowen RK, Sponaugle S (2009) Larval dispersal and marine population connectivity. *Annual review of marine science* **1**, 443-466.

DiBattista J, Berumen M, Gaither M, Rocha L, Eble J, Choat J, MT Craig, Skillings D, Bowen B (2013) After continents divide: comparative phylogeography of reef fishes from the Red Sea and Indian Ocean. *Journal of Biogeography* **40**, 1170-1181.

DiBattista JD, Howard Choat J, Gaither MR, Hobbs JPA, Lozano-Cortés DF, Myers RF, Paulay G, Rocha LA, Toonen RJ, Westneat MW (2016a) On the origin of endemic species in the Red Sea. *Journal of Biogeography* **43**, 13-30.

DiBattista JD, Roberts MB, Bouwmeester J, Bowen BW, Coker DJ, Lozano-Cortés DF, Howard Choat J, Gaither MR, Hobbs JPA, Khalil MT (2016b) A review of contemporary patterns of endemism for shallow water reef fauna in the Red Sea. *Journal of Biogeography* **43**, 423-439.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution* **23**, 347-351.

Fernandez-Silva I, Randall JE, Coleman RR, DiBattista JD, Rocha LA, Reimer JD, Meyer CG, Bowen BW (2015) Yellow tails in the Red Sea: phylogeography of the Indo-Pacific goatfish *Mulloidichthys flavolineatus* reveals isolation in peripheral provinces and cryptic evolutionary lineages. *Journal of Biogeography* **42**, 2402-2413.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**, 489-496.

Gaither MR, Bernal MA, Coleman RR, Bowen BW, Jones SA, Simison WB, Rocha LA (2015) Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Molecular Ecology* **24**, 1543–1557.

Gaither MR, Rocha LA (2013) Origins of species richness in the Indo-Malay-Philippine biodiversity hotspot: evidence for the centre of overlap hypothesis. *Journal of Biogeography* **40**, 1638-1648.

Gralka M, Stiewe F, Farrell F, Moebius W, Waclaw B, Hallatschek O (2016) Allele Surfing Promotes Microbial Adaptation from Standing Variation. *bioRxiv*, 049353.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

Kemp J (1998) Zoogeography of the coral reef fishes of the Socotra Archipelago. *Journal of Biogeography* **25**, 919-933.

Kinlan BP, Gaines SD (2003) Propagule dispersal in marine and terrestrial environments: a community perspective. *Ecology* **84**, 2007-2020.

Kirk H, Freeland JR (2011) Applications and implications of neutral versus non-neutral markers in molecular ecology. *International journal of molecular sciences* **12**, 3966-3988.

Kulbicki M, Parravicini V, Bellwood DR, Arias-Gonzàlez E, Chabanet P, Floeter SR, Friedlander A, McPherson J, Myers RE, Vigliola L (2013) Global biogeography of reef fishes: a hierarchical quantitative delineation of regions. *PloS one* **8**, e81847.

Leis J (1991) The pelagic stage of reef fishes: the larval biology of coral reef fishes. In: *The Ecology of Fishes on Coral Reefs* (ed. Sale PF), pp. 183-230. Academic Press, San Diego, California.

Leray M, Beldade R, Holbrook S, Schmitt R, Planes S, Bernardi G (2010) Allopatric divergence and speciation in a coral reef fish: The three-spot dascyllus, *Dascyllus trimaculatus* species complex. *Evolution* **64**, 1218-1230.

Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Lowe WH, Allendorf FW (2010) What can genetics tell us about population connectivity? *Molecular Ecology* **19**, 3038-3051.

Palumbi SR (2003) Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications* **13**, 146-158.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* **7**, e37135.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Ridgway T, Sampayo EM (2007) Population genetic status of the Western Indian Ocean: what do we Know? *Western Indian Ocean journal of marine science* **4**, 1-10.

Riegl BM, Purkis SJ (2012) Coral reefs of the Gulf: adaptation to climatic extremes in the world's hottest sea. In: *Coral Reefs of the Gulf*, pp. 1-4. Springer.

Rocha L, Bernal, MA, M. R. Gaither, M.E. Alfaro (2013) Massively parallel DNA sequencing: the new frontier in biogeography. *Frontiers of Biogeography*, 67-77.

Thoppil PG, Hogan PJ (2009) On the mechanisms of episodic salinity outflow events in the Strait of Hormuz. *Journal of Physical Oceanography* **39**, 1340-1360.

Victor BC (1991) Settlement Strategies and Biogeography of Reef Fishes. In: *The Ecology of Fishes on Coral Reefs* (ed. Sale PF), pp. 231-260. Academic Press, San Diego, California.

Visram S, Yang M-C, Pillay RM, Said S, Henriksson O, Grahn M, Chen CA (2010) Genetic connectivity and historical demography of the blue barred parrotfish (*Scarus ghobban*) in the western Indian Ocean. *Marine Biology* **157**, 1475-1487.

Waples RS, Punt AE, Cope JM (2008) Integrating genetic data into management of marine resources: how can we do it better? *Fish and Fisheries* **9**, 423-449.

Ward R, Woodwark M, Skibinski D (1994) A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. *Journal of Fish Biology* **44**, 213-232.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Wellington G, Victor B (1989) Planktonic larval duration of one hundred species of Pacific and Atlantic damselfishes (Pomacentridae). *Marine Biology* **101**, 557-567.

CHAPTER 2, SUPPLEMENTARY MATERIALS

SUPPLEMENTARY MATERIALS AND METHODS

Genomic DNA was extracted using the Qiagen DNeasy animal blood and tissue kit (Qiagen, Valencia, USA). Library was prepared using the double-digest RADseq protocol (Peterson *et al.* 2012). I used SphI and MluCl restriction enzymes (New England Biolabs). Pools consisted of 16 barcoded individuals. Size selection of 376-450 bp was carried out using a 2% agarose Pippin Prep® cassette (Sage Science). Unique Illumina Indexes were added to each pool, and the libraries were amplified using a Real-Time PCR library amplification kit (Kapa Biosystems). The concentration of each pool was quantified using a High Sensitivity Kit on a 2100 Bioanalyzer (Agilent Technologies), and then standardized and merged into a single pool. The library was sequenced on a single Illumina HiSeq 2000 lane, at the UCLA Neuroscience Genomics Core facility. I obtained 140 million raw reads (100 bp single end). Raw data was de-multiplexed, trimmed to 95 bp and reads with Phred scores below 10 were discarded, using the "process_rad_tags" script available in STACKS v1.09 (Catchen *et al.* 2013; Catchen *et al.* 2011). Sequencing resulted in 127.7 million quality filtered reads. Individuals with less than 25,000 retained reads were discarded.

Loci were assembled using the STACKS de novo_map.pl pipeline. I chose a minimum of three identical reads to create a stack (m=3), three mismatches allowed between loci within an individual (M=3), five mismatches when aligning reads (N=5), and two mismatches when building the catalog (n=2). Assembly in

STACKS resulted in a total of 535,904 loci. I used the “populations” script to filter loci and create output files, making sure that loci were shared between the seven populations (p=7), in at least 65% of individuals within a group (r=0.65) and with coverage of 8x (m=8). I used only the first SNP of each sequence, and removed loci with minor allele frequencies lower than 1.5% (i.e., at least two individuals must have the unique allele). The filtering resulted in a total of 1,174 loci, and a data matrix 84% complete. For all downstream analyses, I used the STRUCTURE output file produced by STACKS which was converted to other file formats using PGDSPIDER 2.0 (Lischer & Excoffier 2012).

Data analysis

First I identified outlier loci across all populations using the modified FDIST approach (Beaumont & Nichols 1996; Excoffier *et al.* 2009) implemented in ARLEQUIN (EXCOFFIER & LISCHER 2010). The method simulates a distribution of FST VS. expected heterozygosity based on an island model of migration. Outlier loci were those with significantly higher or lower FST than expected by the model (p-value < 0.01). I ran 50,000 simulations with 100 demes per group, with minimum and maximum expected heterozygosities of 0 and 0.5.

Using these results, I classified each locus into three categories: 1) divergent, with FST significantly higher than expected 2) balancing, with FST significantly lower and 3) neutral, all the other loci. After classifying loci, I created three datasets, and applied population genetic analysis to each one separately: 1) neutral loci, n=1,117, 2) outlier loci (only the divergent ones, n=25). 3) All loci together n=1,174 (1,117 neutral, 25 divergent and 32 balancing).

AMOVAs were performed in ARLEQUIN with 10,000 permutations to test for genetic structure between the NWIOP and WIOP provinces. Pairwise FST values (Weir & Cockerham 1984) were calculated between all populations, using 10,000 permutations in ARLEQUIN. A discriminant analysis of principal components (DAPC) (Jombart *et al.* 2010) was executed using ADEGENET (Jombart 2008) for R (R Development Core Team 2015). The DAPC plot represents the individuals as dots and the groups (populations) as inertia ellipses. Ellipse centers are at the gravity center of each population's cloud of points. The plot uses the best number of principal components (PCs) identified with the cross validation method ("xValDapc" function). The "find.clusters algorithm" was also calculated, to determine the number of clusters in the data. In addition, I ran the Bayesian clustering method implemented in STRUCTURE with correlated allele frequencies in an admixture model, one million MCMC and 100,000 burnin chains. The most likely number of clusters (K) was determined with the Evanno method (Evanno *et al.* 2005) using STRUCTURE HARVESTER (Earl & vonHoldt 2012). DISTRUCT (Rosenberg 2004) was used for the graphics in Fig. 2.1. To test for IBD I compared matrixes of FST/(1-FST) and minimum ocean distance, estimated by plugging the coordinates and measuring the closest ocean distances with GOOGLE EARTH 7.1.2 (Google Inc., Mountain View, CA, USA). Mantel tests were performed with GENEPOP (Raymond & Rousset 1995).

SUPPLEMENTARY RESULTS

Global Fst for all the loci was 0.0127 (p<0.0001). Pairwise F_{ST} comparisons indicated significant differences between all populations except Zanzibar with Mayotte or Juan de Nova, and Mayotte with Juan de Nova. An AMOVA showed a very low but significant divergence between the NWIOP (Red Sea, Djibouti and Oman) and the WIOP (Chagos, Zanzibar, Mayotte, Juan de Nova) (F_{CT} =0.0099, p=0.0225).

SUPPLEMENTARY TABLES

Table 2.S1. Pairwise Fst, all loci n=1,174

	NRS	DJI	OMA	DGA	ZAN	MAY	JNO
NRS	****						
DJI	**0.0071**	****					
OMA	**0.0074**	**0.0095**	****				
DGA	**0.0124**	**0.0141**	**0.0231**	****			
ZAN	**0.0107**	**0.0084**	**0.0131**	-0.0013	****		
MAY	**0.0042**	**0.0120**	**0.0153**	**0.0010**	-0.001	****	
JNO	**0.0079**	**0.0136**	**0.0109**	**0.0028**	-0.002	-0.0031	****

Table 2.S2. Outlier loci with significant alignments and e-value below 1e-04 in nucleotide BLAST (Search updated Aug 13, 2016).

Loci	**Description**	**Query cover**	**E-value**	**Identity**
24280	*Cyprinus carpio* genome assembly common carp genome, scaffold 000000138, GenBank: LN590755.1	63%	5e-05	82%
6844	PREDICTED: *Stegastes partitus* vacuolar protein sorting-associated protein 45-like (LOC103376270), mRNA	98%	1e-17	84%
63017	*Epinephelus coioides x Epinephelus lanceolatus* voucher ECEL001 microsatellite ECELXB002 sequence, GenBank: JQ732809.1	83%	2e-09	80%
7990	*Dicentrarchus labrax* chromosome sequence corresponding to linkage group 1, top part, complete sequence, GenBank: FQ310506.3	54%	6e-04	87%
31414	*Cyprinus carpio* genome assembly common carp genome, scaffold: LG32, chromosome: 32, GenBank: LN590696.1	93%	5e-05	74%

Table 2.S3. AMOVA combinations in neutral and divergent outlier loci

Dataset	**Groups defined**	**FCT**	**Variation explained by the groups**	**Probability**
Neutral (n=1,117)	NWIOP and WIOP provinces	0.0041	0.41%	p=**0.0283**
	Red Sea + Djibouti vs. Oman+WIOP	0.0005	0.05%	p=0.4780
	Red Sea + Oman vs. Djibouti+WIOP	0.0035	0.35%	p=0.0879
Divergent outliers (n=25)	NWIOP and WIOP provinces	0.2349	23 %	**p=0.0342**
	Red Sea + Djibouti vs. Oman+WIOP	0.1054	10.54%	p=0.1417
	Red Sea + Oman vs. Djibouti+WIOP	0.1649	16.49%	P=0.1095

SUPPLEMENTARY FIGURES

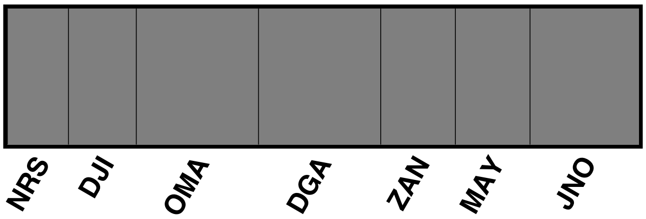

Figure 2.S1. Structure plot, k=1 according to structure harvester.

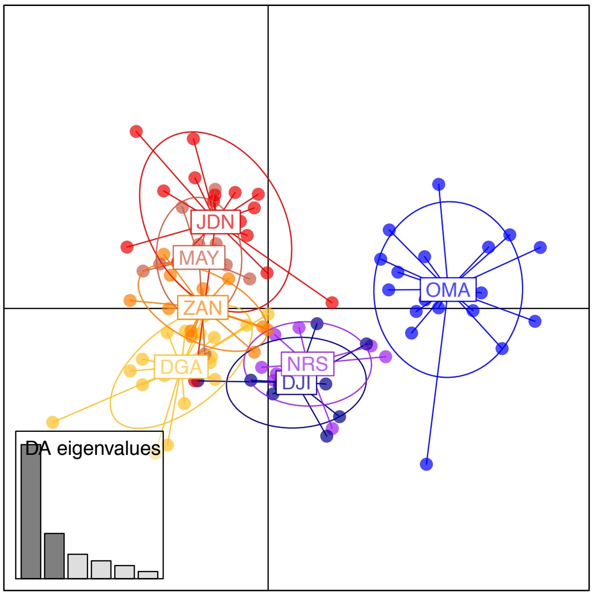

Figure 2.S2. DAPC for all loci, only one cluster was supported.

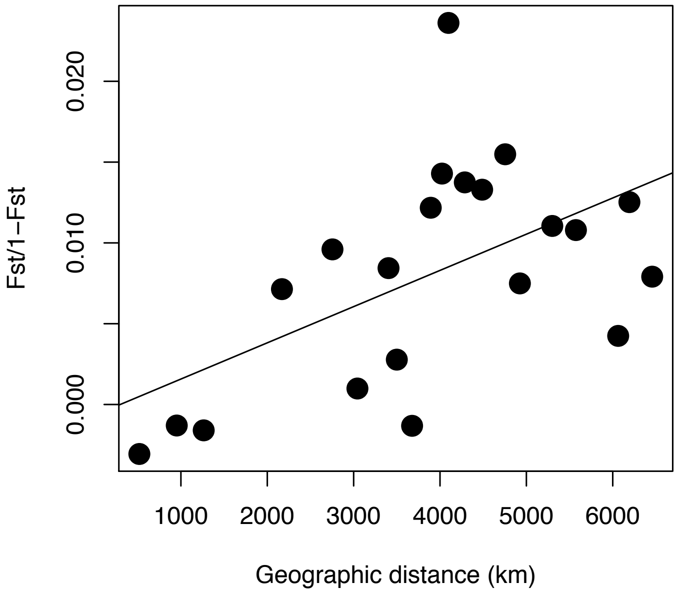

Figure 2.S3. Isolation by distance, r^2=0.3006. Mantel test p=0.0310

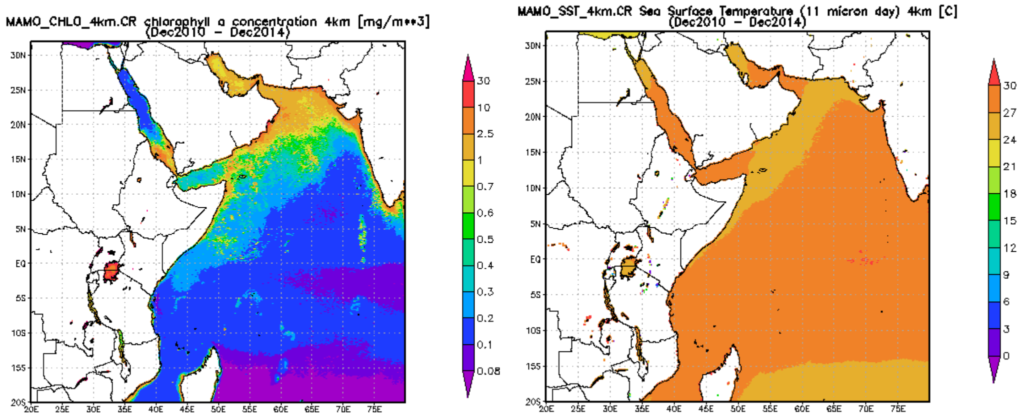

Figure 2.S4. Distribution of average chlorophyll a (left) and sea surface temperature (right) over the study region between 2010 and 2014.

SUPPLEMENTARY REFERENCES

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **263**, 1619-1626.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* **7**, e37135.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Chapter 3

Genomics and subtle color reveal cryptic hybridization in a coral reef fish

CHAPTER 3

GENOMICS AND SUBTLE COLOR REVEAL CRYPTIC HYBRIDIZATION IN A CORAL REEF FISH

ABSTRACT

Cocos (Keeling) and Christmas Islands sit at the border of the Indian and Pacific Ocean bioregions. They represent a secondary contact zone for multiple marine species and populations that were isolated during Pleistocene glaciations. They are also considered a marine suture zone, where closely related species from separate biogeographic regions overlap and interbreed. To date, 15 cases of hybridization between coral reef fishes have been documented in this region, making it an interesting area for speciation studies. In this project, Restriction Site Associated DNA sequencing (RADSeq), indicates hybridization between highly differentiated Pacific and Indian Ocean genetic clades of the three-spot dascyllus, *Dascyllus trimaculatus*. Since there are only subtle differences in color between Indian and Pacific populations, hybridization in this pair had remained undetected. These analyses reveal that the pattern within the suture zone is not homogeneous: Cocos Islands have a genetically stable population that underwent hybridization in the past, whereas hybridization is ongoing at Christmas Island. This project exemplifies how genomic techniques are unraveling more cases of hybridization in marine suture zones, while showing that patterns of introgression can be quite different at relatively small spatial scales. This research is important for

understanding how hybridization contributes to evolutionary novelty of Indo-Pacific coral reefs.

INTRODUCTION

Hybridization, or the interbreeding between species or subpopulations, can have different evolutionary outcomes (Hewitt 1988), providing insight on the dynamics of gene flow and speciation. Hybridization is common in areas of secondary contact where allopatric species meet (Hewitt 1988). Some hybrid zones are located in biogeographic boundaries, in these regions multiple species pairs encounter each other and can hybridize. These are called "suture zones", and include all the individual hybrid zones of interbreeding pairs of taxa (Remington 1968). Hybrid zones are windows on evolution and natural laboratories for investigating the different evolutionary outcomes.

Hybridization has a variety of impacts on the process of speciation. It may slow or reverse differentiation by allowing gene flow and recombination (Abbott *et al.* 2013). With sufficient gene flow, it may remove adaptive gene variants or cause reverse speciation or introgressive extinction (Rudman & Schluter 2016). If hybrids have lower fitness, then it can become an evolutionary dead-end (Barton 2001). In other cases, hybrids may have higher fitness than their parents (hybrid vigor). Hybridization it may inject new genetic variation and it may accelerate speciation via adaptive introgression (Barton 2001). Overall, it may have multiple effects depending on the level of divergence between species (or subpopulations) and population dynamics (abundance, spatial distribution, among others).

There are a few known marine suture zones. Recently it was discovered that Socotra Archipelago is a suture zone for reef fishes, this island sits in the intersection of three biogeographic provinces in the Indian Ocean (DiBattista *et al.* 2015). However the most studied marine suture zone is located in Cocos (Keeling) and Christmas islands (Eastern Indian Ocean). During the glaciations the sea level dropped 130 meters and Sunda Shelf got exposed several times(Ludt & Rocha 2015), causing isolation and speciation. Allopatric species now come into contact near these islands. To date, 15 cases of hybridization between species pairs have been confirmed, involving 27 species across eight families (Hobbs & Allen 2014). Hybridization is more prevalent in the Chaetodontids, followed by Acanthurids and Pomacanthids (Hobbs & Allen 2014). There is an apparent bias between locations, 14 of the 15 hybrid crosses have been identified at Christmas Island, compared to eight present at Cocos (Hobbs & Allen 2014). This bias has been attributed to sampling efforts.

Here I report a new case of hybridization in Cocos (Keeling) and Christmas, involving populations of a Pomacentrid reef fish, *Dascyllus trimaculatus.* This species belongs to a complex that recently diverged from *Dascyllus reticulatus* (another species complex) in the Pleistocene, around 3.9 million years ago (McCafferty *et al.* 2002). *D. trimaculatus* complex has been divided into five mitochondrial clades: 1) the Marquesas endemic *D. strasburgi*, 2) the Hawaiian *D. albisella*, 3) the French Polynesian *D. trimaculatus,* 4) the Pacific rim comprising two introgressed groups: *D. trimaculatus* and *D. auripinnis*, and 5) the Indian Ocean *D. trimaculatus* (Bernardi *et al.* 2003;

Bernardi *et al.* 2001; Bernardi *et al.* 2002; Leray *et al.* 2010; McCafferty *et al.* 2002). Mitochondrial studies put the origin of *Dascyllus trimaculatus* in the Indian Ocean and allopatric speciation explains the different clades, perhaps combined with some parapatric speciation in the range edges eastward.

Here I use a combination of RADSeq data, mitochondrial DNA and field observations to describe the hybridization dynamics of *Dascyllus trimaculatus* clades from the Indian and Pacific oceans. I also describe consistent phenotypic differences that I found among the clades. Furthermore I describe the spatial genetic footprints of this interaction in Cocos and Christmas, where I find ancient and current hybridization. I show that there are subtle differences in coloration between pure Pacific and Indian Ocean *D. trimaculatus* and I used those differences to look at the distribution and abundance of the two clades in the marine suture zone.

METHODS

Sampling

The samples consisted of a selection of genotype classes previously identified from mitochondrial, microsatellite and SNP data. There are strong genetic differences between the Indian and Pacific *Dascyllus trimaculatus* clades, so I selected three populations from each of the pure clades (Indian Ocean and Pacific Ocean) to compare with the hybrid zone individuals (Cocos Keeling and

Christmas) (Fig. 3.1). The “Arabian Peninsula (APE)” population contains 21 individuals from Thuwal, two from Farasan (Saudi Arabia) and three from Socotra (Yemen). The “Indonesia (IND)” population contains three individuals from Manado and four from Komodo. Pooling within these sites didn’t affect the results, preliminary F_{ST} analysis didn’t show significant genetic differences between these samples prior to pooling.

Fish colour morphs and visual surveys

I observed that there were color differences between the two clades, the Indian Ocean *Dascyllus trimaculatus* has a black rear end of the dorsal fin, whereas the Pacific Ocean one has a transparent/white rear end (Fig. 3.2). I also noticed that in Cocos and Christmas there were intermediate individuals with the rear half black and half transparent (or other mixture proportions) (Fig. 3.3). I explored if there were consistent differences in color between Pacific and Indian Ocean populations, using some of the authors’ personal field observations, colleagues’ observations and geo-referenced images from Flickr.

The abundance of *D. trimaculatus* at Christmas and Cocos (Keeling) Islands was estimated using standard underwater visual census methods. Three replicate 50 x 5 m belt transects were conducted at each of two depths (5 and 20 m) across eight sites at Christmas Island, and at seven sites at the Cocos (Keeling) Islands. Individuals were categorised into three phenotypes based on: a) black rear end of the dorsal fin, b) clear rear end of the dorsal fin, c) intermediate forms where about half of the rear was black and the other was transparent.

RADSeq library preparation, sequencing and assembly

Genomic DNA was extracted using the Qiagen DNeasy animal blood and tissue kit (Qiagen, Valencia, USA). DNA concentration was first measured in a fluorometer with the Qubit HS dsDNA essay kit (Life Technologies), and each sample was standardized to 500 ng. The library was prepared using the double-digest RADSeq protocol described by Peterson *et al.* (2012). Samples were digested for 3 hours at 37°C, using restriction enzymes SphI and MluCl (New England Biolabs). Digests were quantified with Qubit HS dsDNA essay kit (Life Technologies) and a Qubit 2.0 Fluorometer, and then cleaned with Dynabeads M-270 Streptavidin (Life Technologies). Ligation was performed with P2 universal adaptors. Additionally, sets of 16 individuals were barcoded with unique P1 adaptors. After ligation, each group of 16 individuals was pooled and bead cleaned. Pools were size-selected for a range of 400 bp using a 2% agarose gel that was ran for 45 minutes, and purified with the Zymoclean Gel DNA Recovery Kit. Unique Illumina Indexes were added to each pool, and the libraries were amplified with 10 PCR cycles, using the high fidelity Platinum Taq DNA polymerase (Thermo Fisher Scientific). The concentration of each pool was quantified using a High Sensitivity Kit on a 2100 Bioanalyzer (Agilent Technologies), and then standardized and merged into a single pool for sequencing. Samples were sequenced in two lanes of an Illumina Hi-Seq 2000, at

KAUST Genomics Core facilities, which resulted in 251,425,358 and 270,346,963 single end raw reads (100bp).

Loci were assembled using STACKS v1.09 (Catchen *et al.* 2013; Catchen *et al.* 2011). Raw data was demultiplexed and filtered using the "process_rad_tags" script. Average quality scores were determined within a sliding window 15% the length of the sequence, and reads with any score below 90% of being correct were discarded. Sequences were trimmed to 95bp and loci were assembled using the STACKS "de_novo_map.pl" pipeline, using a minimum of three identical reads to create a stack (m=3), three mismatches allowed between loci within an individual (M=3), five mismatches when aligning reads (N=5), and two mismatches when building the catalog (n=2). I used the "populations" script to obtain loci shared between the eight populations (p=7), in at least 65% of individuals within a population (r=0.65) and with coverage of 8x (m=8). I used only the first SNP of each sequence, and removed loci with minor allele frequencies lower than 5%. The filtering resulted in a total of 128 individuals with 2,818 loci and a data matrix 83% complete (see Fig. 3.1 for samples per site). For all downstream analyses, I used the STRUCTURE output file produced by STACKS which was converted to other file formats using PGDSPIDER 2.0 (Lischer & Excoffier 2012).

RADSeq data analysis

To quantify the extent of genetic differentiation, pairwise F_{ST} values were estimated between Cocos-Keeling, Christmas and the Pacific/Indian Ocean clades using ARLEQUIN 3.5.1.2 (EXCOFFIER & LISCHER 2010), using 10,000 permutations. Genetic assignment of individuals was calculated with STRUCTURE (Pritchard *et al.* 2000) using correlated allele frequencies in an admixture model, one million Markov chain Monte Carlo (MCMC) repetitions and 100,000 burn-in runs. I ran 10 simulations for each K (from 1 to 9). The most likely number of clusters (K) was determined with the Evanno method (Evanno *et al.* 2005) using STRUCTURE HARVESTER (Earl & vonHoldt 2012). DISTRUCT (Rosenberg 2004) was used to generate the graphical display of population structure. To further investigate population structure, I ran a discriminant analysis of principal components (DAPC) (Jombart *et al.* 2010). This multivariate method seeks to show differences between groups as best as possible while minimizing variation within populations. It does not rely on any particular population genetics model, uses the information of all the loci, and generates a graphical representation of the relatedness between the populations (Jombart *et al.* 2010). The analysis was executed using ADEGENET (Jombart 2008) for R (R Development Core Team 2015), using the best number of principal components (PCs) identified with the cross validation method ("xValDapc" function). The "find.clusters algorithm" was also calculated, to investigate the number of genetic groupings in the data.

Mitocondrial DNA sequencing

A 495 base pair fragment of the mitochondrial control region (D-loop) was amplified in 121 individuals using CR-A and CR-E primers (Lee *et al.* 1995) (See Fig 3.1. for sample size per site). PCR reactions (10 µl) were performed using 5µl of multiplex PCR mix (Qiagen), 0.5 µl of CRA 10µM, 0.5 µl of CRE 10µM, 3 µl of water, 1 µl of DNA (30-100ng/µl). Touchdown PCR reactions were set up as follows: 15 min. at 95°C, followed by 20 cycles of 30s at 95°C, 60 s at 58°C, 90s at 72°C. During these cycles the temperature was decreased -0.4°C every minute. This was followed by 15 cycles of 30s at 95°C, 60s at 50°C, 90s at 72°C; and a final extension of 72°C for 10 min. DNA was purified using exonuclease I and FastAP^{TM} thermosensitive alkaline fosfatase (ExoFAP; USB, Cleveland, OH, USA), running it for 60 min at 37°C, followed by 15 min at 85°C. DNA was sequenced in the forward and reverse direction using fluorescent-labeled dye (BigDye 3.1, Applied Biosystems Inc., Foster City, CA, USA) using an ABI 3730xl analyzer. Sequences were aligned and trimmed in Geneious 6.06 (Drummond *et al.* 2013) and final edits were performed by eye.

Mitocondrial DNA data analysis

To generate the haplotype networks and the neighbor joining tree, I determined the best model of sequence evolution in JMODELTEST2 (Darriba *et al.* 2012). Since the model which identified by the Bayesian information criterion was not available in ARLEQUIN, I selected the second most similar which was the Kimura 2P. I also analyzed the haplotype and nucleotide diversity using ARLEQUIN. To assign each individual haplotype from the hybrid zone to the Indian and Pacific clades, I constructed a neighbor joining tree using PAUP (Swofford 2003). Also, a minimum spanning haplotype network was constructed using the software POPART (Leigh & Bryant 2015). Typically, reef fish hybridization studies identify hybrids by their intermediate phenotype, and then try to determine how they assign to the parental mitochondrial clades. I didn't have phenotype information for all of them, and the intermediate phenotype is difficult to detect so I determined which of the Cocos and Christmas individuals were hybrids and which ones were "pure" based on the SNP genotype found by STRUCTURE, and complemented that information with any available phenotype data. This information allowed me to compare the mitochondrial assignment with the SNP data assignment and the color morph when possible.

RESULTS

Fish colour morph and visual survey results

I was able to identify that the Indian and Pacific Ocean adult *Dascyllus trimaculatus* clades have consistent phenotypic differences. The Indian Ocean clade consistently has a dark rear end of the dorsal fin, while the Pacific Ocean one has a clear (or white) rear (Fig. 3.2). To date, all the observed individuals and documented pictures were consistent with this pattern. The observations came from the Red Sea (Saudi Arabia, Egypt, Jordan), from the Indian Ocean (Oman, Somaliland, Kenya, Madagascar, Chagos, Maldives); from the Pacific Ocean: Okinawa, American Samoa, Taiwan (Dongsha atoll, Green Island), Micronesia (Ulithi, Yap, Pohnpei, Kosrae), Fiji, Guam, Phillipines, Malaysia, Sulawesi (over 150 observations), Bali, Thailand, Vanuatu, South Wales, Australia and French Polynesia. To date, I have only found both morphs in Christmas, Cocos-Keeling and Bali, Indonesia. In Cocos and Christmas, I identified individuals with intermediate morphotypes: the rear edge of the dorsal fin in some cases was only partially transparent/white.

Underwater surveys of *D. trimaculatus* at Christmas Island revealed that the clear fin individuals were the most common phenotype and their density (0.961 per 250m2 +/- 0.29 SE) was approximately double that of black fin individuals (0.451 per 250m2 +/- 0.15 SE). In contrast, at the Cocos Islands the black fin individuals (0.381 per 250m2 +/- 0.14 SE) were the most common phenotype of *D. trimaculatus* and the clear fin individuals were rare (0 per 250m2). Individuals with half clear/black were rare at both Christmas (0.098 per 250m2 +/- 0.066 SE) and Cocos Islands (0 per 250m2). Although clear fin individuals, or half clear/black individuals were not observed in transects at the Cocos Islands, they were occasionally seen outside of transects.

RADSeq results

According to F_{ST} comparisons, there is strong genetic differentiation between the Pacific and Indian Ocean clades, and between the Pacific and Cocos (Table 3.1). All comparisons were significant, but Pacific vs. Christmas, and Indian vs. Cocos had the least differentiation of all comparisons (Table 3.1). STRUCTURE results and the Evanno method showed that the most likely number of clusters is two (k=2) (Fig 3.4). In Christmas, the sampled population was composed of pure Pacific, pure Indian and admixed genotypes. The analysis identified F1 hybrids and backcrosses. In Cocos I found a different genetic composition. All the individuals were backcrosses with > 80% of their genotype composition assigning to the Indian Ocean. I didn't find pure Pacific, pure Indian or F1 hybrid genotypes. The DAPC analysis (Fig. 3.5) strongly supports the findings of the STRUCTURE assignment. The Pacific and Indian Ocean clades are separated. Individuals from Christmas have a range of genotypes, most of them matching the character of the Pacific ones, a lesser proportion matching the Indian genotypes, and in the middle are individuals of Christmas with intermediate genotypes. On the other hand, the analysis suggests that individuals from Cocos-Keeling are more closely related to the Indian pure populations, and some of them match intermediate Christmas genotypes, but overall the population stands out as its own.

Mitochondrial DNA results

The haplotype and nucleotide diversity are shown in Table 3.2. Neighbor joining tree results showed that Cocos haplotypes only group with the Indian Ocean mitochondrial clade, whereas Christmas haplotypes fell into both Indian and Pacific clades, with 7 and 15 individuals respectively. The haplotype network (Fig. 3.6) shows the distribution of Christmas haplotypes in relation to their SNP genotypes and color morphs. The Cocos individuals with mitochondrial data were classified either as unknown SNP genotype or backcrossed. All of the sampled Cocos individuals had a black rear end of the dorsal fin, suggesting that they had an Indian Ocean phenotype. Christmas individuals were classified into pure Pacific, pure Indian, F1, backcrossed SNP genotypes, and one individual with pure Pacific SNP genotype but discordant coloration. Unfortunately I didn’t have phenotype information for all the Christmas individuals that were genotyped, but I had pictures of four individuals: the F1 SNP genotype had intermediate phenotype: half of its rear fin was transparent. A backcross with a dominant Indian SNP genotype had an Indian phenotype. A backcross with a dominant Pacific SNP genotype had a Pacific phenotype. However, another backcross with a dominant Pacific SNP genotype had an Indian phenotype (with a black rear of the dorsal fin), its coloration was discordant with its genotype.

The haplotype network results shows there are clear differences between the Pacific and Indian clades, all the individuals collected from Pacific locations grouped together, and all the individuals collected from Indian locations grouped together. The individuals from the hybrid zone had a different pattern: All the Cocos individuals assigned to the Indian clade, but the Christmas assigned to either clade. At Christmas, the pure Indian SNP genotypes fell into either Pacific or Indian mitochondrial clades. This is interesting because in locations outside of the hybrid zone, there is concordance between the mitochondrial and SNP genotypes: the SNP and mitochondrial genotypes always assign to the same clade, either Indian or Pacific. The pure pacific SNP genotypes from Christmas fell into the pacific mitochondrial clade, even the individual with discordant coloration (rear of dorsal fin was black but its SNP genotype was almost pure pacific). The hybrid F1 fell also into the Pacific clade. The Pacific backcrosses fell into either

Pacific or Indian clade, and the Indian backcross fell into the Pacific clade. These results suggest that there is ongoing hybridization and introgression in Christmas, and that there is bidirectional maternal contribution, because backcrosses are found on both mitochondrial clades. It also suggests that individuals are so deeply backcrossed in Christmas, that pure genotypes do not assign to their expected mitochondrial clades or their expected coloration, as they do outside of the hybrid zone.

DISCUSSION

Cryptic hybridization

This is the first report of *Dascyllus trimaculatus* clades hybridizing at Cocos and Christmas. The two clades have strong genetic differentiation and will be soon described as separate species (Salas*, in prep*.). Leray *et al.* (2010) predicted that hybridization of *D. trimaculatus* would be found in these islands, and the result is not unexpected. However, these hybrids are cryptic. Most of the hybridization cases described before at Cocos and Christmas used color as the main character to identify hybrids and then tested the hybridization hypothesis with genetics (Hobbs & Allen 2014). If I had only used genetic data from the mitochondrial DNA, hybridization would have remained undetected. There are subtle differences in color between pure Pacific and pure Indian morphs, but it is difficult to detect intermediate (hybrid) morphs. So this study stresses the importance of SNPs in detecting cryptic hybridization. SNPs may also be useful to identify backcrosses of more conspicuous hybrid fish (like *Chaetodon* and *Centropyge*) and understand the relationship of hybrids and color, which as I show, may not be directly related. There may be more species pairs that hybridize at Cocos and Christmas but that are cryptic (Hobbs & Allen 2014). There are a number of species that have breaks between the Pacific and Indian Ocean (Gaither & Rocha 2013), many of those may be hybridizing as well.

This is the nature of this hybridization case study: It seems to be influenced by the unequal densities of parents, maternal contribution is bidirectional, there is bi-directional introgression, and the parents do not seem to have assortative mating, but more ecological observations need to be done to confirm that. Color data is scarce, but my preliminary data of color and genetics suggest that at some point the color of the backcrosses is not indicative of its genetic profile.

Hybrids at Christmas

The fish surveys and genetic data support that hybridization is ongoing at Christmas. SNP data demonstrates the presence of both parent clades, F1 and a mixture of backcrosses in the Bayesian population assignment. Independent analysis (DAPC), add support to these results by showing the composition of the Christmas population, with individuals more closely related to parental populations and also the presence of intermediate ones. Mitochondrial data and fish surveys add support to the observation of ongoing hybridization, since both Pacific and Indian haplotypes are found, and fish surveys indicate the presence of Pacific, Indian and intermediate color morphs. I also observed fish of both Pacific and Indian Ocean phenotypes mating in Christmas (Fig 3.2).

At Christmas, pure Pacific Ocean fish are more common than pure Indian Ocean fish. This may be due to its closer geographic location to the Pacific and the direction of prevailing currents coming from the Pacific. Christmas island closest location is Java, Indonesia, at approximately 350 km (James & McAllan 2014). I know that individuals sampled in Komodo Indonesia are from the Pacific clade (this study), and I also know that parts of western Australia (Ningaloo) are from the Pacific clade (personal unpublished data). I also know that there are both morphs in Bali, Indonesia (this study). Indian ocean larvae may be arriving occasionally at Christmas via stepping stone dispersal from Indian Ocean locations like Chagos, Sri Lanka, Myanmar or Northern Indonesia (i.e. Sumatra). It's possible that the break of the clades is consistent with the marine's Wallace line (Barber *et al.* 2000), and the break sits around Java or Bali. The prevailing currents allow larvae from both Pacific and Indian Oceans to arrive to Christmas. The South Java Current (SJC) crosses the coast of Sumatra and diverges to the West near Java, bringing water from the Indian Ocean. The South Equatorial Current (SEC) and the Indonesian Through Flow (ITF) come from the East bringing water from the Pacific Ocean (Yang *et al.* 2015).

Hybridization of *Dascyllus trimaculatus* at Christmas may be neutral or evolutionary advantageous, but it does not seem detrimental. I observed increased *Dascyllus trimaculatus* gene diversity. The Pacific mitochondrial clade of pure Pacific *Dascyllus trimaculatus* populations is characterized by lower genetic diversity than the Indian Ocean one (Leray *et al.* 2010). However in Christmas, individuals with Pacific haplotypes have higher diversity than pure Pacific populations. So it seems that the hybridization with Indian Ocean *Dascyllus trimaculatus* is increasing genetic diversity of Pacific haplotypes at Christmas. Considering the population as a whole, their nucleotide diversity is larger than anywhere else (Table 3.2). The populations at Christmas due to the ongoing hybridization may have an evolutionary advantage by having a larger diversity pool than other places and genetic material could be available for evolutionary innovation.

Hybrids at Cocos

The nuclear and mitochondrial data results are consistent with the idea that hybridization happened in the past but is no longer ongoing or is largely infrequent at Cocos. There may not be frequent arrival of pure Pacific larvae. Neither is a normal pure Indian Ocean population. All the individuals that were sampled were backcrosses with ~80% of their genotype of the Indian Ocean, showing introgression of Pacific Ocean genotype in their genomes. All the individuals genotyped would have been classified as pure Indian Ocean morphs, if

I didn't had the SNP data. So again, here I stress the importance of using SNPs to uncover cryptic introgression of the genome.

Cocos (Keeling) is more isolated than Christmas, and also farther away from the Pacific Ocean. It is located ~900 km from Christmas and more than 1,000 km from Indonesia or any other land (Hobbs *et al.* 2014). The fish surveys indicate that pure pacific morphs and intermediate forms are rare, none was observed on the surveys, but clear fin individuals have been found outside of surveys. Cocos may be receiving occasional vagrants from the Pacific Ocean, perhaps not enough to maintain a genomic signature of ongoing hybridization. Its populations may or may not be isolated from the Indian Ocean. The SNP data indicates no presence of pure Indian Ocean individuals so I tend to favor the idea that Cocos is really isolated from both Pacific and Indian Ocean populations.

I observed decreasing *Dascyllus trimaculatus* gene diversity at Cocos (Table 3.2), relative to the Indian Ocean clade. There is no evidence of a general bottleneck; its diversity is larger relative to the Pacific Ocean clade. It may be the product of hybridizing with the less diverse Pacific clade, or it may be lower than the rest of the Indian Ocean due to its geographic isolation. The populations at Cocos may have an evolutionary disadvantage by having a lower diversity pool than other places and less genetic material could be available for evolutionary innovation.

Geographic mosaic of hybridization

One of the most striking results of this study is that in a hybrid zone there can be heterogeneous patterns of hybridization that depend on the geographic mosaic of dispersal and gene flow of pure populations. Cocos and Christmas are only separated by 900 km and the genetic distinction is really striking. There are very large genetic differences within populations of a single species and also between hybrid populations at a rather small scale. These patterns have very important ecological and evolutionary implications. The differences are led by population dynamics, such as relative abundances of pure types, genetic characteristics of these pure types and dynamics of population connectivity. The genetic structure of Cocos fish has changed to a point that it has become a divergent population. In the future it may become more divergent or alternatively the pacific fingerprint may eventually disappear. The introgression may have changed its evolutionary outcome injecting novel genotypes with unknown fitness consequences. This could be a mechanism for parapatric speciation in its very early stages.

Hobbs (2014) has studied hybridization of many species in Cocos and Christmas, and he reports an apparent bias of hybrids between locations, Christmas has 14 hybrid crosses, Cocos has 8 hybrid crosses only. The authors attributed that bias to sampling intensity. However it may be due to the isolation of Christmas and may be causing different evolutionary outcomes in those species and many others at these two geographic locations, possibly changing the evolutionary trajectory of a whole community.

TABLES

Table 3.1. *Dascyllus trimaculatus* pairwise F_{ST} (2,818 SNP loci). All comparisons were significant

	Indian	**Cocos**	**Christmas**	**Pacific**
Indian	***			
Cocos	0.0241	***		
Christmas	0.1284	0.0815	***	
Pacific	0.2526	0.1827	0.0330	***

Table 3.2. *D. trimaculatus* haplotype and nucleotide diversity. Hybrid zone individuals were separated according to their haplotypes (Pacific or Indian).

	Indian Haplotype		**Pacific Haplotype**	
	Haplotype diversity	Nucleotide diversity, PI	Haplotype diversity	Nucleotide diversity, PI
Indian pure populations	1 (31)	2.2% (31)	-	-
Pacific pure populations	-	-	0.99 (34)	1.0% (34)
Cocos population	0.99 (34)	1.5% (34)	-	-
Christmas population	1 (7)	2.1% (7)	1 (15)	1.9% (15)
Christmas (pooled, n=22)	Haplotype diversity: 1+/-0.0137, Nucleotide diversity 3.9%+/-2%			

FIGURES

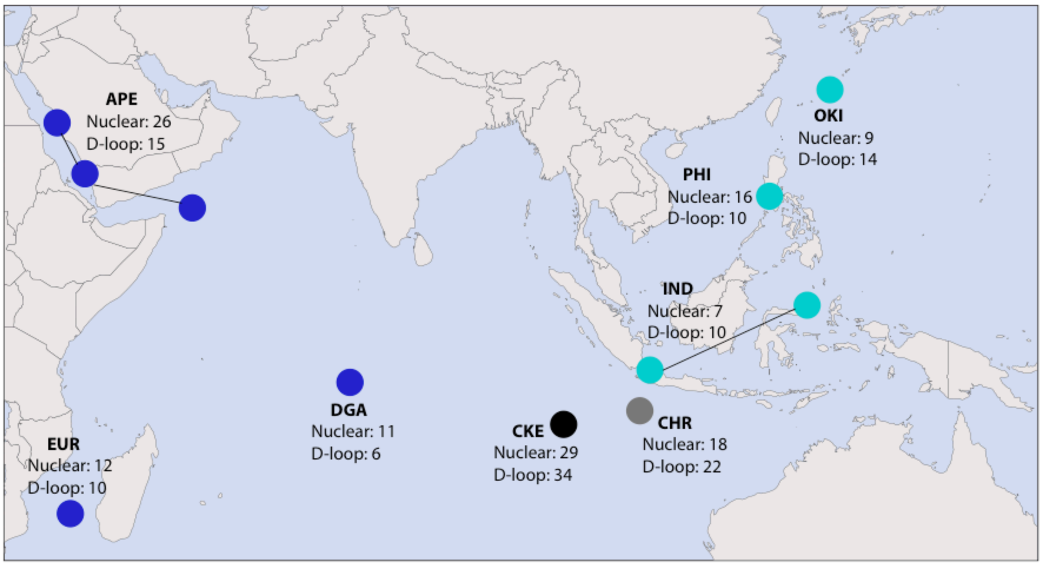

Figure 3.1. Sampling locations and sample sizes. Indian Ocean populations: APE: Arabian Peninsula; EUR: Europa, Scattered Islands; DGA: Diego Garcia, Chagos Archipelago; Hybrid zone populations: CKE: Cocos-Keeling; CHR: Christmas; Pacific populations: IND: Indonesia; PHI: Philippines; OKI: Okinawa.

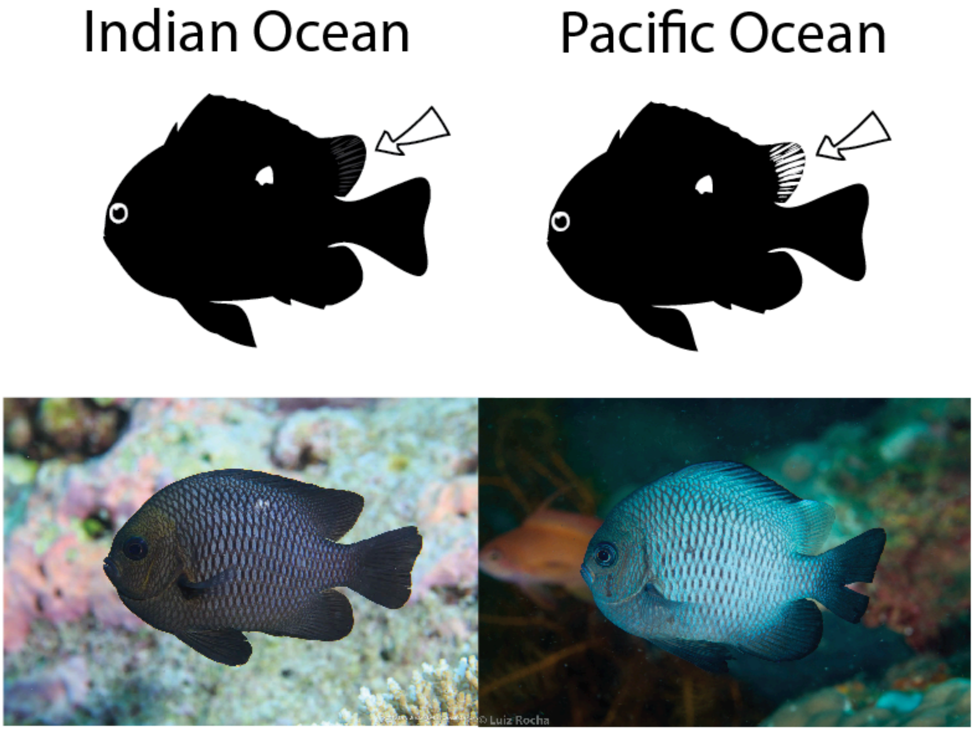

Figure 3.2. Indian and Pacific Ocean *Dascyllus trimaculatus*. The morphs are recognized by the coloration of the rear end of the dorsal fin. The left picture was taken in the Maldives (photo by Tane Synclair-Taylor), and the right picture was taken in the Philippines (photo by Luiz A. Rocha). The white coloration in the body is variable; it changes to darker depending on the behavioral display of the fish and it can happen in both the Indian and Pacific Ocean clades.

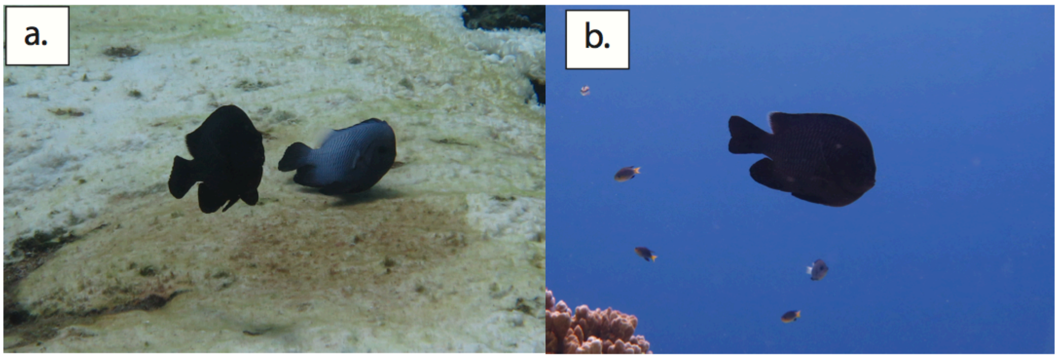

Figure 3.3. (a) Christmas is one of the few places where Pacific and Indian morphs overlap and interbreed. (b) I have identified fish with various ranges of intermediate coloration on the rear of the dorsal fin, but the intermediate coloration is very subtle.

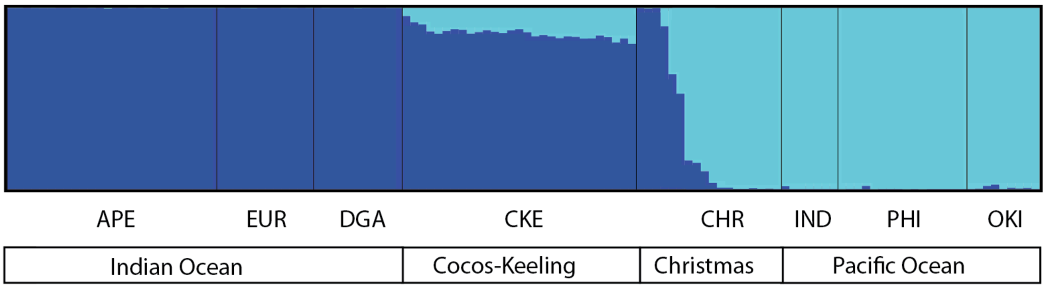

Figure 3.4. Bayesian cluster analysis performed by STRUCTURE based in 2,818 SNP loci. The most likely number of clusters (k) was two (Delta K=14,566), based on the Evanno method.

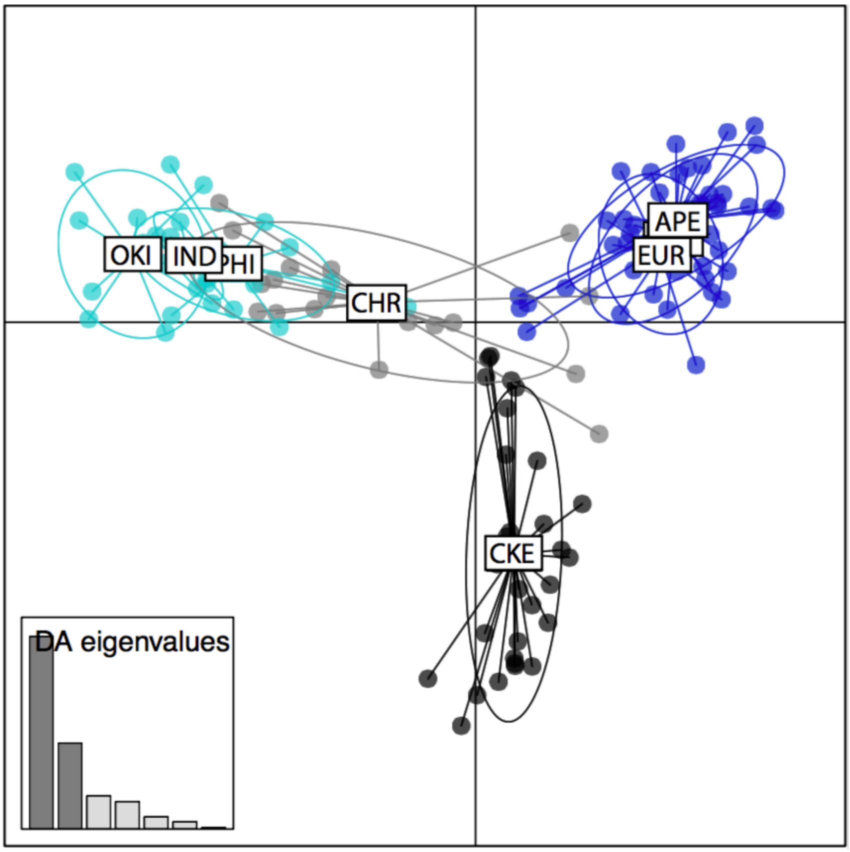

Figure 3.5. Discriminant Analysis of Principal Components (DAPC) based on 2,818 SNP loci, showing the relationships of Christmas (CHR, grey) and Cocos-Keeling (CKE, black) *Dascyllus trimaculatus* with pure Pacific (OKI, IND, PHI, light blue) and Indian Ocean (APE, EUR, DGA, dark blue) populations.

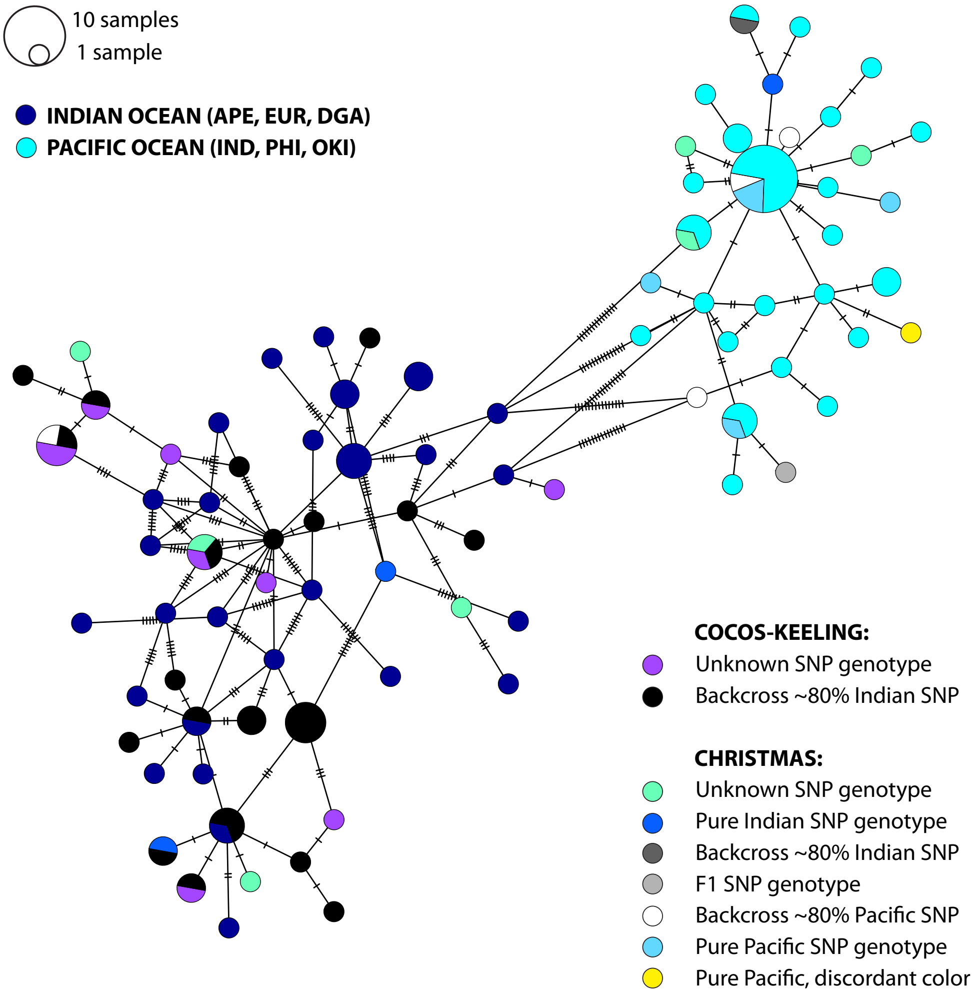
10 samples
1 sample
INDIAN OCEAN (APE, EUR, DGA)
PACIFIC OCEAN (IND, PHI, OKI)
COCOS-KEELING:
Unknown SNP genotype
Backcross ~80% Indian SNP
CHRISTMAS:
Unknown SNP genotype
Pure Indian SNP genotype
Backcross ~80% Indian SNP
F1 SNP genotype
Backcross ~80% Pacific SNP
Pure Pacific SNP genotype
Pure Pacific, discordant color

Figure 3.6. Haplotype network (minimum spanning network), using 495 bp of the mitochondrial D-loop. Circles represent haplotypes, with its size proportional to the haplotype frequency. Connecting lines and each bar represent single mutation steps. Cocos-Keeling and Christmas individuals were classified based on the SNP genotypes into: F1 hybrids, backcrosses, pure Pacific or Indian Ocean genotype. Individuals without SNP data were classified as "unknown". There was one individual collected at Christmas Island with an almost pure Pacific SNP genotype, but the color of the rear of the dorsal fin was black, the typical color of the Indian Ocean phenotype. This individual was called here "Pure Pacific, discordant color".

REFERENCES

Abbott R, Albach D, Ansell S, Arntzen J, Baird S, Bierne N, Boughman J, Brelsford A, Buerkle C, Buggs R (2013) Hybridization and speciation. *Journal of Evolutionary Biology* **26**, 229-246.

Barber PH, Palumbi SR, Erdmann MV, Moosa MK (2000) Biogeography: a marine Wallace's line? *Nature* **406**, 692-693.

Barton N (2001) The role of hybridization in evolution. *Molecular Ecology* **10**, 551-568.

Bernardi G, Holbrook S, Schmitt R, Crane N (2003) Genetic evidence for two distinct clades in a French Polynesian population of the coral reef three-spot damselfish *Dascyllus trimaculatus*. *Marine Biology* **143**, 485-490.

Bernardi G, Holbrook SJ, Schmitt RJ (2001) Gene flow at three spatial scales in a coral reef fish, the three-spot dascyllus, *Dascyllus trimaculatus*. *Marine Biology* **138**, 457-465.

Bernardi G, Holbrook SJ, Schmitt RJ, Crane NL, DeMartini E (2002) Species boundaries, populations and colour morphs in the coral reef three–spot damselfish (*Dascyllus trimaculatus*) species complex. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **269**, 599-605.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772-772.

DiBattista JD, Rocha LA, Hobbs JPA, He S, Priest MA, Sinclair-Taylor TH, Bowen BW, Berumen ML (2015) When biogeographical provinces collide: hybridization of reef fishes at the crossroads of marine biogeographical provinces in the Arabian Sea. *Journal of Biogeography* **42**, 1601-1614.

Drummond A, Ashton B, Cheung M (2013) Geneious v6. 0. Available from http://www.geneious.com.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Gaither M, Rocha LA (2013) Origins of species richness in the Indo-Malay-Phillipine biodiversity hotspot: evidence for the centre of overlap hypothesis. *Journal of Biogeography* **40**, 1638-1648.

Hewitt GM (1988) Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology & Evolution* **3**, 158-167.

Hobbs J-PA, Allen GR (2014) Hybridisation among coral reef fishes at Christmas Island and the Cocos (Keeling) Islands. *Raffles Bulletin of Zoology, Supplement* **30**, 220-226.

Hobbs J-PA, Newman SJ, Mitsopoulos GE, Travers MJ, Skepper CL, Gilligan JJ, Allen GR, Choat HJ, Ayling AM (2014) Fishes of the Cocos (Keeling) Islands: new records, community composition and biogeographic significance. *Raffles Bulletin of Zoology*, 203-219.

James DJ, McAllan IA (2014) The birds of Christmas Island, Indian ocean: A review. *Australian Field Ornithology* **31**, S1.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

Lee W-J, Conroy J, Howell WH, Kocher TD (1995) Structure and evolution of teleost mitochondrial control regions. *Journal of Molecular Evolution* **41**, 54-66.

Leigh JW, Bryant D (2015) popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**, 1110-1116.

Leray M, Beldade R, Holbrook S, Schmitt R, Planes S, Bernardi G (2010) Allopatric divergence and speciation in a coral reef fish: The three-spot dascyllus, *Dascyllus trimaculatus* species complex. *Evolution* **64**, 1218-1230.

Lischer H, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Ludt WB, Rocha LA (2015) Shifting seas: the impacts of Pleistocene sea-level fluctuations on the evolution of tropical marine taxa. *Journal of Biogeography* **42**, 25-38.

McCafferty S, Bermingham E, Quenouille B, Planes S, Hoelzer G, Asoh K (2002) Historical biogeography and molecular systematics of the Indo-Pacific genus *Dascyllus* (Teleostei: Pomacentridae). *Molecular Ecology* **11**, 1377-1392.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* **7**, e37135.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945.

Remington CL (1968) Suture-zones of hybrid interaction between recently joined biotas. In: *Evolutionary biology*, pp. 321-428. Springer.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Rudman SM, Schluter D (2016) Ecological impacts of reverse speciation in threespine stickleback. *Current Biology* **26**, 490-495.

Swofford D (2003) PAUP* ver 4.0. b10. *Phylogenetic Analysis Using Parsimony and Other Methods. Sunderland, MA: Sinauer Associates, Sunderland.*

Yang G, Yu W, Yuan Y, Zhao X, Wang F, Chen G, Liu L, Duan Y (2015) Characteristics, vertical structures, and heat/salt transports of mesoscale eddies in the southeastern tropical Indian Ocean. *Journal of Geophysical Research: Oceans* **120**, 6733-6750.

General conclusions

The goal of my dissertation was to study population genomics of *Dascyllus trimaculatus* in the Indian Ocean at different spatial scales, using RAD markers. Assessing the genetic structure on common and widespread marine species has always been challenging, due to their large effective population sizes and low genetic structure. I studied the patterns of genetic differentiation of this damselfish in the Indian Ocean and in an area of secondary contact between Indian and Pacific Ocean faunas. I chose to work in the Indian Ocean because when I started the dissertation there were few studies on population genetics of marine species within that region. It is a challenging place to work at, because there is limited access to many locations, due to war and piracy.

According to previous research with mitochondrial markers, *D. trimaculatus* in the Indian Ocean had low genetic structure and high genetic diversity. By using thousands of SNP markers, I was able to show that *D. trimaculatus* populations are not panmictic within the Indian Ocean. However, the patterns are cryptic. By using RADSeq, I was able to uncover patterns of genetic structure that would have remained hidden otherwise. I used a novel approach with outlier loci to uncover patterns concordant with phylogeographic breaks present in other marine species, but absent in neutral markers in *D. trimaculatus*.

As I was doing my research, other population genetic studies within the region were conducted and my studies contributed to the common patterns identified among many species. One of those is the strong genetic break found in the southern Red Sea. A steep environmental gradient in salinity, temperature and productivity is reflected in a genetic break found in many species, including *D. trimaculatus* (Chapter 1). The fact that it was found in the loci outliers suggests that divergent selection may be acting on the species and allowing it to adapt to different environmental conditions. I also found genetic differences between the Red Sea and the Arabian Peninsula (Chapter 1), and between the Arabian Peninsula and the Western Indian Ocean (Chapter 2). These differences may be attributed to the complex history of the Red Sea and Arabian Peninsula, the lack of suitable habitat on the coasts of Oman and Somalia, strong upwellings and the changing environmental conditions found in the Red Sea and Arabian Peninsula.

The Indian Ocean *D. trimaculatus* has strong genetic differences with the Pacific Ocean populations. In Chapter 3, I demonstrated that the SNP markers can also be useful to discover cryptic hybridization. I found that *Dascyllus trimaculatus* Pacific and Indian Ocean genetic clades meet in Cocos-Keeling Islands and Christmas Island and interbreed. There is hybridization and backcrossing that was only discovered after using thousands of SNP markers. I also found that there are consistent differences between the Pacific and Indian Ocean clades in the color of the rear of the dorsal fin. Future work will separate *D. trimaculatus* into two species.